SOCIAL USER MINING: USER PROFILING OF SOCIAL MEDIA NETWORK BASED ON MULTIMEDIA DATA MINING

Mohammed Ali Eltaher

Under the Supervision of Dr. Jeongkyu Lee

DISSERTATION
SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIRMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOHPY IN COMPUTER SCIENCE

AND ENGINEERING

THE SCHOOL OF ENGINEERING

UNIVERSITY OF BRIDGEPORT

CONNECTICUT

April, 2015

SOCIAL USER MINING: USER PROFILING OF SOCIAL MEDIA NETWORK BASED ON MULTIMEDIA DATA MINING

APPROVALS

Committee Members

Name	Signature	Date
Dr. Jeongkyu Lee		4/21/15
Dr. Julius Dichter		4,28.15
Dr. Miad Faezipour		4,22,15
Dr. Navarun Gupta		4/21/15
Dr. M. Emre Celebi		4/11/15

Ph.D. Program Coordinator

Dr. Khaled M. Elleithy — 5/7/15

Chairman, Computer Science and Engineering Department

Dr. Ausif Mahmood — 5/7/2015

Dean, School of Engineering

Dr. Tarek M. Sobh — 5/7/2015

SOCIAL USER MINING: USER PROFILING OF SOCIAL MEDIA NETWORK BASED ON MULTIMEDIA DATA MINING

SOCIAL USER MINING: USER PROFILING OF SOCIAL MEDIA NETWORK BASED ON MULTIMEDIA DATA MINING

ABSTRACT

In recent years, the pervasive use of social media has generated extraordinary amounts of data that has started to gain an increasing amount of attention. Each social media source utilizes different data types such as textual and visual. For example, Twitter is used to transmit short text messages, whereas Flickr is used to convey images and videos. Moreover, Facebook uses all of these data types. From the social media users' standpoint, it is highly desirable to find patterns from different data formats.

The result of the huge amount of data from different sources or types has provided many opportunities for researchers in the fields of data mining and data analytics. Not only the methods and tools to organize and manage such data have become extremely important, but also methods and tools to discover hidden knowledge from such data, which can be used for a variety of applications. For example, the mining of a user's profile on social media could help to discover any missing information, including the user's location or gender information. However, the task of developing such methods and tools is very challenging. Social media data is

unstructured and different from traditional data because of its privacy settings, data noise, and large capacity of data. Moreover, combining image features and text information annotated by users reveals interesting properties of social user mining, and serves as a useful tool for discovering unknown information about the users. Minimal research has been conducted on the combination of image and text data for social user mining.

To address these challenges and to discover unknown information about users, we proposed a novel mining framework for social user mining that includes: 1) a data assemble module for different media source, 2) a data integration module, and 3) mining applications. First, we introduced a data assemble module in order to process both the textual and the visual information from different media sources, and evaluated the appropriate multimedia features for social user mining. Then, we proposed a new data integration method in order to integrate the textual and the visual data. Unlike the previous approaches that used a content based approach to merge multiple types of features, our main approach is based on image semantics through a semi-automatic image tagging system. Lastly, we presented two different application as an example of social user mining, gender classification and user location.

DEDICATION

To the memory of my father, Ibrahim and my brother, Mohammed, whom inspired me and gave me the talent in engineering and science.

ACKNOWLEDGEMENTS

First and foremost, my thanks are wholly devoted to Allah who has helped me all the way to complete this work successfully. Then, I would like to express my deepest appreciation to my advisor Dr. Jeongkyu Lee for giving me the opportunity to be creative. He encouraged me at difficult times and offered valuable suggestions when I faced tough challenges. This dissertation would not have been possible without his supervision.

Second, to my committee members, Dr. Julius Dichter, Dr. Miad Fazipour, Dr. Navarun Gupta, and Dr. M. Emre. Thank you for your encouraging and constructive feedback. Reviewing a thesis is never an easy task, and I am grateful for their valuable and insightful comments.

Third, to the faculty and staff of School of Engineering at University of Bridgeport. I am proud to be a member of this family. Thank you for helping me to develop the skills I need to complete this thesis.

Last but not least, this dissertation would not have been possible without the love and support of my family: Mom, wife, kids, brothers, and sisters. Thank you all for your encouragement.

Table of Contents

ABSTRACT iv

DEDICATION vi

ACKNOWLEDGEMENTS vii

LIST OF TABLES xi

LIST OF FIGURES xii

CHAPTER 1: INRTODUCTION 2

1.1 Overview 2

1.2 Research Problem 4

1.3 Research Scope 7

1.4 Motivation Behind the Research 7

1.5 Contributions of Proposed Research 8

1.6 Structure of Dissertation 9

CHAPTER 2: LITERATURE SURVEY 10

2.1 Content Based Social User Mining 10

2.1.1 Textual Information 12

2.1.2 Visual Information 16

2.1.3 Textual and Visual Information 18

2.2 Semantic Based Social User Mining 19

2.2.1 Image Semantic 21

2.2.2 Video Semantic 22

2.3 Data Mining Techniques in Social User Mining 23

2.3.1 Supervised Approach 25

2.3.2 Unsupervised Approach 27
2.4 Discussion 28
CHAPTER 3: MINING FRAMEWORK 30
3.1 Overview 30
3.2 Framework Structure 30
3.3 Discussion 31
CHAPTER 4: DATA ASSEMBLE 32
4.1 Overview 32
4.2 Data Collection 33
4.2.1 Ground Truth Data 34
4.2.2 Textual and Visual Data 36
4.2.2 Semantic Data 37
4.3 Preprocessing 39
4.4 Representation 40
4.4.1 Textual Data 40
4.4.2 Visual data 40
4.4.3 Semantic Data 41
4.4.4 Discussion 41
CHAPTER 5: DATA INTEGRATION 43
5.1 Overview 43
5.2 Content Based Image Fusion 44
5.2.1 Integrated Data Units 44
5.2.2 Integration Scheme 45
5.3 Semantic Based Image Fusion 46
5.3.1 Integrated Data Units 47

5.3.2 Integration scheme 48
5.4 Conclusion 49
CHAPTER 6: MINING APPLICATION 50
6.1 Gender Classification 50
6.1.1 Overview 50
6.1.2 Classification Algorithms 50
6.1.3 Content Based Classification 52
6.1.4 Semantic Based Classification 54
6.1.5 Experiments & Discussion 55
6.2 User Location 59
6.2.1 Overview 59
6.2.2 Gaussian Mixture Model (GMM) 59
6.2.3 Modeling Location 60
6.2.4 Local Word Selection 62
6.2.5 Experiments & Discussion 63
CHAPTER 7: CONCLUSIONS AND FUTURE WORK 68
7.1 Research Contributions 68
7.2 Publications 69
7.2.1 Journal 69
7.2.2 Conference 69
7.2.3 Relevant Poster Award 70
7.3 Future work 70
REFERENCES 72

LIST OF TABLES

Table 2.1	Content based social user mining	11
Table 2.2	Semantic based social user mining	20
Table 2.3	Data mining techniques in social user mining	24
Table 4.1	Ground truth data set	35
Table 6.1	Tag gender dictionary	53
Table 6.2	Example of Hue histogram	53
Table 6.3	Example of Hue in bag of words	54
Table 6.4	Experiment result for content based classification	57
Table 6.5	Experiment result for semantic based classification	58
Table 6.6	The top-30 frequency resorted local words (GMM, NL)	65
Table 6.7	The baseline experiments result	66
Table 6.8	Location Estimation Result	66
Table 6.9	Example of correctly estimated cities and corresponding tweet massages	67

LIST OF FIGURES

Figure 1.1	General diagram of social user mining	6
Figure 3.1	An overview of mining framework	31
Figure 4.1	Data assemble module	33
Figure 4.2	An example of Flickrs' user profile	34
Figure 4.3	Crawler for user profile	35
Figure 4.4	Example about textual and visual data which are represented by tags and images of user	37
Figure 4.5	A semi-automatic image tagging system (Akiwi)	38
Figure 4.6	Example of semantic data collection	39
Figure 5.1	Contents based data integration module	44
Figure 5.2	Feature vector of the content based data integration	46
Figure 5.3	Semantic based data integration module	47
Figure 5.4	Feature vector of the semantic based data integration	48

CHAPTER 1: INRTODUCTION

1.1 Overview

With the overwhelming popularity of social media websites, we have witnessed the generation of massive data on the web, which offer a new opportunity to resolve some challenges in multimedia mining. The amount of social media data such as photos, videos, and text has grown at an extraordinary rate. According to [1], as of 2009, Flickr users have shared over 4 billion images and videos on the site, Facebook users have shared a similar amount of photos each month, and YouTube users have shared 20 new hours of video content every minute. Moreover, in late 2010, Twitter had 175 million registered users worldwide and they produced 65 million tweets per day [2]. In fact, social media sites such as Flickr and Twitter have provided different platforms for a user to share different types of information.

The simplicity of social media has generated a huge amount of information, which may appear in different ways. Some social media sites have a platform that can support textual information, while others support visual information or both textual and visual. For example, Flickr users can perform different activities such as posting photos and marking favorite photos. Mining these activities is needed in order to build

meaningful applications. Furthermore, social media presents significant opportunity for the development of many applications and services [3].

Although social media data is different from traditional data types, data mining techniques can be used for the purpose of mining user information from social media data. Data mining is the process of extracting interesting patterns or knowledge from huge amounts of data [4, 5]. This information can be applied to evolving social media topic such as community detection, clustering, statistical analysis, classification and association rules mining. In order to cover a variety of these data mining research topics, many of the top algorithms such as SVM, Naive Bayes, k-NN were developed [6].

Even though some researchers have already applied the data mining techniques in order to study different problems with in social media mining, mining user-generated data is still very challenging due to its unstructured, large size, and noisy data [7]. In addition, due to privacy concerns, social media users tend to hide some of their profile information [8]. For social media service providers, users' profile information is useful in order for the providers to customize their services to the users in many ways such as recommendations and customer relationship management (CRM).

In this dissertation, we investigate the problem of how to utilize data mining techniques in order to understand social media data. In general, social media mining is the process of representing, analyzing, and extracting actionable patterns of social media data [9]. We tackle the challenges in mining social media data for the purpose of discovering hidden information about users and exploiting user-generated data for different applications such as gender or location prediction.

1.2 Research Problem

Social user mining is the process used to discover unknown (or hidden) information about users from their publicly-available data on social media. We focused on different types of information such as gender, ethnicity, age, and marital status. Some user information is intentionally blocked by a user because of privacy issues, while others are unknown patterns. It is possible to mine temporal and spatial patterns. For example, in the case of home location, a moving pattern would be a good candidate. Most of the publicly-available data on social media is user-generated content. We can define social user mining as follows:

Definition 1 (*Social media data*): A social media data d is a multimedia document created by a social media network user. d Consists of multimedia data contents, including text, audio, image, and video.

The social media data d is a basic item of social user mining. Example of d can be viewed as a tweet on Twitter[1] or a photo with meta-data in Flickr[2]. Generally, social media data is based on a variety of multimedia data types such as text, audio, image, and video.

Definition 2 (*Social user*): A social user $u = \{a_1, \dots, a_n\}$ is a user who creates social media data d_u, where a_n is n number of user attributes describing user profile information.

[1] http://www.twitter.com
[2] http://www.flickr.com

For example, a Flickr user u_f has profile information that includes user-id, location, name, gender, and marital status. In addition to the user profile, there might be unknown user attributes, such as moving patterns which are not part of profile information.

Definition 3 (*Social user mining*): Given a set of social media data $D_u = \{d_1, \dots, d_k\}$ that is relevant to a user u, social user mining is a process to discover one or more missing attributes of u, i.e., $a_1, \dots, a_k$ are attributes of u.

In the social media networks, there are many attributes to be considered for mining purpose, which are related to the user such as gender, marital status, location, and age. In this dissertation, gender and location attributes have been utilized. Social user mining is still difficult because of the following reasons:

1. In order to make good structured data from different media sources, data preprocessing is required.
2. It is important to identify the subset of content that have higher level concepts and are most useful to discover user information. Social media sites have different types of data, such as user-generated contents, social network features, and meta-data. For example, Flickrs' users can upload photos to their account. In addition, they can select some photos as their favorite photos. Thus, if we consider the gender attribute for mining, indeed selecting the favorite photos will be beneficial compare to selecting the general photos of user.

3. Select the effective and scalable data mining techniques to mine such social user information. Different social media data require different techniques to build any application.

Moreover, Figure 1.1 summarizes an overview of the social user mining process that includes different media sources, social media data for social users, data preprocessing and the social user mining process.

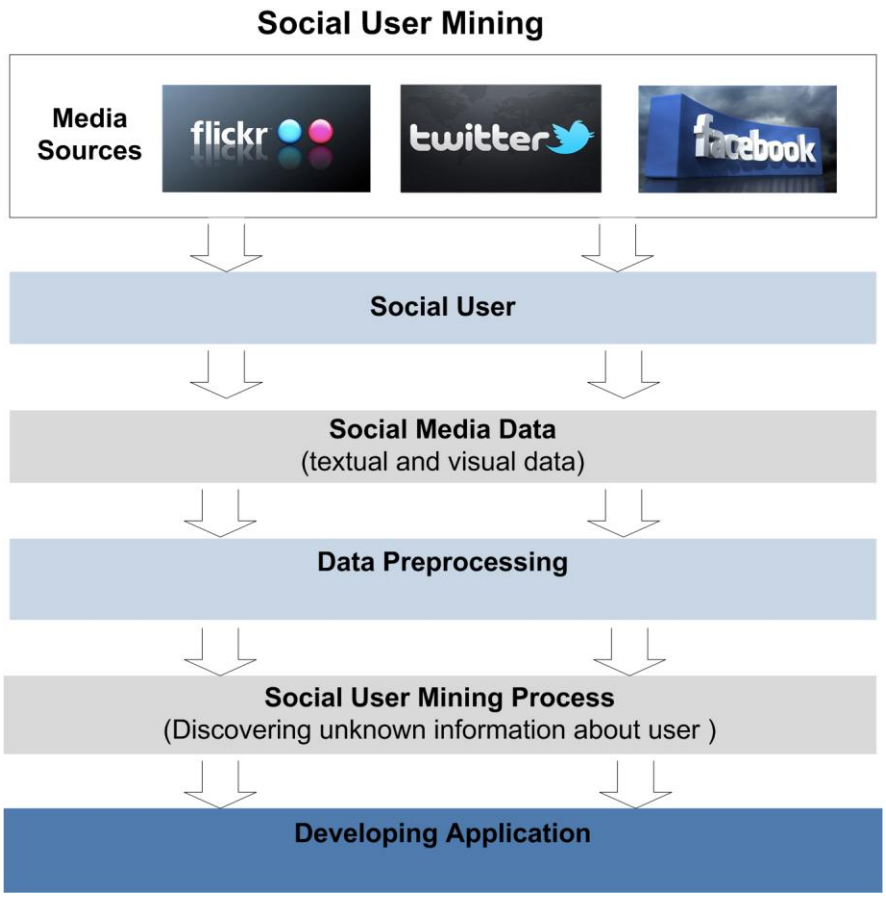

Figure 1.1 General digram of social user mining, which describe the process of mining any application with social media data including: social media sources, users, type of data, data preprocessing, mining process, and application development.

1.3 Research Scope

An important observation of social networks is that they contain a rich level of information about users such as shown in user's tag, comment, descriptive text, and image. This information can be used in order to enhance the social user mining process. For data sources, we focused on Flickr, one of world's best online photo management and sharing applications, and Twitter, used to share information daily. For data modalities, we will use users' tags and tweets as textual information, and users' photos as visual information.

Previous studies on social user mining do not provide us with the proper methods to manage different data modalities. In this research, we present a new method for data integration applied to both text and image for social user mining. Our research objectives are to listed below:

- Provide the proper methods in order to understand and manage social user data.
- Discover unknown (or hidden) information about users from their publicly-available data found on social media.
- Evaluate the appropriate multimedia features for the purpose of social user mining.

1.4 Motivation Behind the Research

The explosive growth of social media networks over the Internet has led to a massive volume of information available on the web. Millions of users in different social

media sources connect to each other, express themselves, and share interests through the web [10]. With a growing number of users in social media, being able to discover hidden information about every user becomes important to many applications. For example, mining user demographic information, such as gender, has a potential to extract actionable patterns that can be useful for business, users, and consumers. If businesses have access to users' demographics, such as gender, such information can be useful in targeted online marketing.

Although some researchers have already studied different problems found in social media mining, available approaches that apply text and image data for social multimedia mining are limited. Our idea is motivated by the fact that using different modalities of data can plays a key role to enhance the social user mining research.

1.5 Contributions of Proposed Research

To best of our knowledge, current researches on social user mining does not provide the proper methods to understand and manage social user data. This dissertation will bridge this gap and contribute to discover unknown information about users based on multimedia data mining. We summarize the potential contributions of this dissertation as follows:

1. A novel mining framework is proposed for social user mining that contains a data assemble module for different media sources, a data integration module, and mining applications.

2. A data assembly module is introduced in order to process both textual and visual information from different media sources.
3. A new data integration method is developed in order to integrate textual and visual data based on image semantic through a semi-automatic image tagging system.
4. Two different mining applications are presented as examples of social user mining, gender classification, and user location.

1.6 Structure of Dissertation

The rest of this dissertation is organized as follows. In Chapter 2, we investigate the related work of mining social media data. More specifically, the survey focuses on three aspects: (1) contents based social user mining, such as textual, visual, and both textual and visual information; (2) semantics based social user mining; and (3) social user mining based on mining techniques. Chapter 3 presents our novel mining framework for social media users. In Chapter 4, we establish a data module for a social user to model both textual and visual data. There are three main tasks associated with this module: (1) data collection, (2) data preprocessing, and (3) data representation. Chapter 5 presents a novel approach of information fusion by utilizing tags and images of users to enhance the social user mining. We use two different approaches: (1) contents based image fusion and (2) semantic based image fusion. In Chapter 6, we introduce two different mining applications as examples of social user mining with a specific focus on gender classification and user location problems. Finally, Chapter 7 concludes our dissertation, and presents the future research directions.

CHAPTER 2: LITERATURE SURVEY

Recently, research in social media mining has been reviewed by [1, 11]. An approach for mining social multimedia data based on a number of application highlighted by [1]. Naaman et al. focused on two social media applications, Flickr Landmark, and Concert Sync. Another social mining approach by [11] has presented a comprehensive study of the state of the art in social media analysis. The study is based on three aspects, graph-theoretic approaches, applications of semantic web technologies, and data mining and analytic. This chapter provides a comprehensive survey on recent research on social user mining. In particular, the survey focuses on three aspects: (1) contents based social user mining, such as textual, visual, and both textual and visual information, (2) semantics based social user mining, and (3) social user mining based on mining techniques.

2.1 Content Based Social User Mining

In recent years, there has been huge expansion of user generated content data in social networks such as Twitter, YouTube, and Flickr. These networks are used daily by millions of users with the goal of sharing and/or consume content from various subjects such as art, politics and economics [12]. In this section, we survey research works on social user mining from the following aspects; (1) social user mining based on textual

information, (2) social user mining based on visual information, and (3) social user mining based on both textual and visual information. An overview of ommon data involves in social user mining are summarized in Table 2.1. Specifically, Table 2.1 summarizes the literatures by three data types: textual, visual information, and both. The data unit is the kind of information for any data source. Tweets in Twitter, wall posts in Facebook and image contents in Flickr are examples of data unit.

Table 2.1 Content based social user mining

Data Type	**Data Source**	**Problem**	**Data unit**
Textual	Twitter	Predict a user location [13, 14, 15]	Tweets
		Characterize real-world event [19]	
		Gender prediction [24]	Tweets, profile
	Facebook	Discover the social emotion [22]	Wall posts, comments
	Netlog	Prediction of age and gender [23]	Chat massages
	Flickr	Events identification on social media [20, 57]	Image context
Visual	Flickr	Estimate geographic information [27]	Image content
		Group Suggest [28]	
Visual and Textual	Flickr	geo-location of image [31]	Image content and context
		Understand user contributed media collection [33]	
		Determine where photo is taken [34]	

2.1.1 Textual Information

Mining the textual information on the social media has been an active research topic in the recent years. The textual data in social user mining can be coming from many different kinds of modalities. For example, an incomplete list of them include tweets in Twitter, user profile with different text fields, such as name, location, and description. There are several user profile information that have been focused in these researches, which can be categorized into one of the followings.

First, the authors of [13, 14, 15, 16] studied the content of the tweets to predict a user location. Twitter is an online social network based on a short text message. In some cases, a user's tweets provide location information, i.e., a name of place directly, but also certain words or phrases to be associated with a specific location. Then, such location from text-based social media data can be predicted by applying data mining techniques.

Z. Cheng et al. [13] focus on the estimation of city level location for a Twitter user. Using a set of tweets from a set of users, they estimate a user's probability of being located in cities across USA. Their approach relies on two key points: a classification component for automatically identifying words in tweets with strong local geo-scope and a lattice-based neighborhood smoothing model for refining a user's location estimate. As results, their location estimator can place 51% of Twitter users within 100 miles of their actual location. However, their approach didn't cover any relationship between different tweet messages such as reply tweet messages.

In order to increase the accuracy, Chandra et al. [14] use the relationship between tweets and reply tweets to estimate a user's location. They evaluate the user's location based on his tweet content along with the content of the related reply tweet messages. Although they were able to improve the accuracy by 10% comparing to the state of the art estimation, further improvements are required to obtain more accurate result. For example, finding more relationship between the user's information and combining them together may help more to estimate the user location in the social network. The authors in [15] further improved the prediction quality of a Twitter user's home location by estimating the spatial word usage probability with Gaussian Mixture models. To remove noises effectively, they also proposed unsupervised measurements to rank the local words. Their approach can achieve a better performance to the state-of-the-art. Furthermore, [16] propose a unified discriminative influence model to solve the problem of profiling users' home locations in the context of Twitter. Their method integrates locations observed from user's friends, followers, and tweets. The experiments shows that their method improve the state-of-the-art methods by 13%.

Second, the textual information can be used for mining events. An approaches that detects events from Flickr images has been proposed by [17, 18]. The authors in [17] use the available user-contributed tags to capture the content of images. They rely on the metadata of time and location to analyze the distribution of images through tags and a wavelet transform is employed to suppress noise. In particular, tags related to events are identified and clustered. Afterwards, for each tag cluster, images corresponding to the represented event are extracted. Another approach by [18] creates a linear combination of

tags, titles, and image descriptions as textual features to identify the event for the images and applies the naïve classifier to these features. Meanwhile, support vector machine is applied to the temporal information. Then, a linear combination of the two classification results is used to identify the events for new images.

In order to determine a relatedness of real-world events, Lee [19] applied unsupervised and supervised learning into relatedness analysis. They have utilized Twitter streams to establishing two relatedness evaluation techniques: a model for online evaluation of emerging events and measure metrics for offline events. Dealing with user generated content in micro-blogs, the study found a challenging language issue in messages that in the informal English domain. However, identifying event relatedness on social media is very important and requires more techniques to identify more events. Another approach for identifying events on social media has been presented by [20]. Given a set of social media documents, the goal was to identify events that are reflected in the documents, as well as the documents that correspond to each event. They use some user provided features such as title, description, and tags as well as some other automated generated features. However, identifying events over social media sites is a challenging problem because social media data are unstructured and noisy.

Third, mining textual information has been used to discover the social emotions on social media networks [21, 22]. Using the wall posts and comments on Facebook, the aim is to identify if the text contain emotion or not [22]. One of the challenging in dealing with text mining on social media is the unstructured language of social media. To overcome this problem, Yassine at el. [22] developed new lexicons that cover common

expressions used by users on social media, social contractions, and Arabic expressions transliterated into English. On the other hand, Bao et al. [21] presented the way of discovering and modeling the connection between online documents and user-generated social emotions. They proposed a new joint emotion topic model for social text mining. At the beginning, the model generates a set of latent topics from emotions, after that it generates affective terms from each topic. Even though their experimental result shows that the model can effectively discover meaningful latent topics and distinguish the topics with strong emotions from background topic, still they should evaluate the model with a larger scale of online document because the collected documents that they use are too small.

Last, mining demographics information such as gender, ethnicity, age, and marital status is an interesting topic for the researcher. As an example, Peersman et al. [23] applied a text categorization approach for the prediction of age and gender on a corpus of chat texts. Their study investigate the automatic prediction of age and gender using short chat messages from Netlog[3]. Moreover, [24, 25] have also presented studies for mining demographic information using Twitter data. Burger et al. [24] investigated the development of high performance classifiers for identifying the gender of Twitter users using content of the tweet text as well as three fields from the Twitter user profile: full name, screen name, and description. The authors in [25] presented a stacked-SVM based classification experiments about different user attributes such as gender, age, regional

[3] http://www.netlog.com

origin, and political orientation. In addition, [26] addressed the task of predicting the gender of the YouTube users based on comments and profile.

2.1.2 Visual Information

Visual information in social media is mainly represented by images or videos. Generally, images have the following features: color, shapes, edges, textures, and regions. However, the selection of the right features is very helpful to discover hidden knowledge about a social user. One of our main goals of this research is to discover the most promising features specifically for social user mining. Based on the image features that we mentioned above, the authors in [27, 28, 29, 30] studied different aspects of mining visual information.

First, Hays et al. [27] proposed a purely data-driven scene matching approach to estimate geographic information from photos. Six image features have been evaluated for this task: tiny images, color histograms, texton histograms, line features, gist descriptor color and geometric context. Their study shows that using all features together performs better than just by itself. On the other hand, if the features are used independently, the most geographically discriminative features are the gist, color histogram, and texton histogram. Although their approach is the first to be able to extract geographic information from a single image, lots of work need to be done for the better results by examining more image features.

Second, the authors in [30] explored a study about discriminating visual preferences on Flickr data. Given a set of preferred images of a Flickr user, they extract a

pool of low and high level features to explore aspects that allow to distinguish different users. The features that they extracted from an image are: color, edges, textures, regions, objects, faces, and scenes. Their test shows that the low level features, i.e., color, texture and regions play a primary role as discriminative features rather than the high level features, i.e., objects and scenes. However, their approach is just a composition of feature extraction and feature weighting. Additionally, [28] proposed to automatically suggest photo groups based on image content. As visual descriptors, they extract four features: tiny image, color histogram, GIST descriptor, and color and edge directive descriptor. As a result, the incorporation of both color and edge histogram information gives the best performance among the other features.

Last, Cheng et al. [29] conducted a study on various multimedia features to obtain intuitive understanding of their impact on AD click behavior in display advertising. Specifically, they developed image and flash features that describe the visual perception of the content of display ADs. To improve the accuracy, they developed a feature selection algorithm to remove the redundant and highly correlated features.

Mining visual information for social user is an important problem and could be useful for many tasks such as location estimation or gender classification. At the same time, it is very important to extract the best features from images that helps to examine and exploit the correlation between image properties and any related task. Based on our survey, we found that color is one of the most widely used visual features and has been extensively studied on the multimedia mining research.

2.1.3 Textual and Visual Information

Textual and visual information are rich sources of social user mining. It is highly desired to integrate one media with another for better accuracy. For example, in order to determine the gender of a user, a profile image and its description can be more effective than a single media such as image itself or description only. Recently, to get such a benefit of using textual and visual information, there has been relatively new study on social user mining as follows.

First, research on georeferenced social media has explored the benefit of employing both textual and visual information. Gallagher et al. [31] use both textual and visual information to find geographical locations of Flickr images. Their model builds upon the fact that visual content and user tags of a picture can be useful to find the geo-location. Visual content is characterized using four features including tiny image, color histogram, GIST descriptor and texton histogram. The combination of visual content matching and local tag probability maps outperform baseline approaches based just on tags or visual contents. Moreover, in a work published in [32], a comprehensive survey on recent research and application on online geo-referenced media was presented. Their study shows how researches used both textual and visual information based on four aspects: organizing and browsing georeferenced media, mining semantic/social knowledge from georeferenced media, learning landmarks in the world, and estimating geographic location of a photo.

Second, Kennedy et al. [33] have combined tag-based location and place information with image content-analysis to improve automated understanding of a large

user contributed media collection on Flickr. Given a set of photographs, they first determine a set of tags that are likely to represent landmarks and geographic features as well as geographic areas where these tags are prominent. Then, they extract visual features from the images that correspond to each tag in its respective areas. With respect to textual information, they identify tags that have events or place semantics. On the other hand, the aspect of visual content was the use of complementary features to capture color, texture, and local point characteristics of images. Despite the fact that their study shows the use of visual information can increase the precision of automatically generated summaries of representative views of locations, the impact of visual information on discovering unknown knowledge such as location or event need further exploration. For example, the way how we handle the visual information is very important because each visual feature is required different process of representation.

Last, the authors in [34] used classification methods for predicting location of photos based on visual, textual and temporal features. They use both textual tags and visual features of photo to determine where it is taken. Their study shows that visual and textual features have different strengths and weaknesses. For example, the visual features have the advantage that they are inherent to the image itself, while the textual tags are available only if a user has added them. However, it is more challenging to automatically find and interpret visual features comparing to textual features.

2.2 Semantic Based Social User Mining

In recent years, social networks for multimedia sharing such as Flickr have become more popular by allowing people to easily upload, share and annotate multimedia

objects with keywords. Labeling the semantic content of multimedia objects such as images with a set of keywords is known as image tagging. More detailed about different types of image tagging can be found in [35]. The social user mining task is mainly depends on the availability and quality of tags. However, the existing studies show that tags are few, impressive, ambiguous, and overly personalize [36]. In addition, recent studies reveal that users do annotate their photos with the motivation to make them better accessible to the general public [37]. A semi-automatic tagging process, that helps to tag a multimedia objects would improve the quality of tagging and thus the overall social user mining process. In general, the aim of an automated multimedia object tagging task is to assign set of semantic keywords to image or video. Table 2.2. Shows an overview of the semantic based user mining related work.

Table 2.2. Semantic based social user mining

Multimedia Data	**Task**	**Methods**
Images	Automated photo tagging	Distance metric learning [38, 39]
		Training of tagged images [40, 41]
		Probabilistic formulation [42]
		Sparse coding representation [43]
	Tag refinement	Graph-based semi-supervised learning[44]
Video	Semi-automatic video tagging	Social knowledge and visual similarity[45]
		Segmentation of cuts and analysis of visual content [46]
		Content redundancy [47, 48]

2.2.1 Image Semantic

In [38, 39], the authors proposed a distance metric learning techniques to automated photo tagging tasks based on Flickr's images. The author in [38] Presented a Probabilistic distance metric learning techniques (PDML). First, they discovered probabilistic side information from the data using a graphical model approach, and then present an effective probabilistic RCA algorithm to find an optimal metric from the probabilistic side information. In the other hand, the authors in [39] proposed unified distance metric learning (UDML) method, which learns metrics from implicit side information hidden in massive social images on the web.

Based on a training set of tagged images, many models have been proposed to associate visual features with semantic concepts keywords. The authors in [40] proposed an annotation method which establishes the correlations between semantic concepts and low-level features. Using a local multi-label classification indictor function, their technique captures the keyword contextual correlations and also exploits the discrimination between visual similar concepts. The authors in [41] proposed a new image tagging framework, which directly takes the noisy social user-tagged images as the training data for learning the reliable image classifiers. In particular, they invent a tag refinement module for identifying and eliminating the noisy tags by substantially exploring and preserving the low-rank nature of the tag matrix and the structured sparse property of the tag errors. Their experiments on two real-world social image databases

illustrate the superiority of their approach as compared to existing methods. On the other hand, their framework need scalability improvement to handle big multimedia data.

In addition, the authors in [42] introduced a probabilistic formulation for semantic image annotation. To addresses the limitations of unsupervised labeling, they presented a Supervised Multiclass Labeling (SML) by explicitly making the elements of the semantic vocabulary the classes of a multiclass labeling problem. Moreover, Zhang et al. [43] proposed a semi-automatic image annotation model based on a sparse coding representation of the images. In order to remove the semantically irrelevant images, their method uses a label transfer mechanism to automatically recommend promising tags to each image by assigning each image a category label first. Based on the results, the recommended keywords can effectively reflect the image content. The authors in [44] applied graph-based semi-supervised learning technique to learn a tag ranking model to achieve tag refinement. They exploit the problem of inferring images' semantic concepts from Flickrs' images and their associated noisy tags.

2.2.2 Video Semantic

In addition to the image tagging, research in video tagging has also received some attention in the latest years. A method for video tag suggestion based on social knowledge and visual similarity has presented by [45]. Specifically, their algorithm suggests new tags that can be associated to a given key-frame using the tags associated to videos and images uploaded to social sites such as YouTube and Flickr and visual features. However, the features used to evaluate the visual similarity of key-frames need further improvement in order to get better suggestion of tags. Moreover, a semi-automatic

tagging system of videos has presented by [46]. This system is based on segmentation of cuts and analysis of visual content of the video. By exploiting visual features extracted from key frames, the tagging module assigns semantic concepts to videos. However, further improvement needed in order to provide a quality assessment of the precision of tags.

Content redundancy in social network environments such as Youtube can be seen as a feature rather than a problem. Study by [47] shows evidence of a significant amount of redundancy can provide useful information about YouTube video. They have used content overlap in the YouTube's video to establish new correlation between videos forming a basis for automatic tagging methods. In addition, the authors in [48] show that content-based links in YouTube videos can provide useful information for generating new tag. They developed two methods of tag assignment by utilizing the video overlap relationships. The first method is called a neighbor-based, which take just immediately overlapping videos. The second method is based on propagation of tag weights within the visual overlap graph. In conclusion, the two studies have shown that content redundancy in social networks can be used to obtain richer annotations for multimedia objects.

2.3 Data Mining Techniques in Social User Mining

There are various mining techniques that can be used in evolving social user mining. In this section, we classify data mining techniques used in social user mining into supervised and unsupervised algorithms. The common example of the supervised approach is classification, whereas the clustering is one of the unsupervised approaches. First, we survey the social user mining based on supervised approach, and then we

investigate the social user mining based on unsupervised approach. Table 2.3 summarizes the mining techniques involved in social user mining based on supervised and unsupervised approaches.

Table 2.3. Techniques based social user mining

Mining Approach	**Techniques**	**Task**
Supervised	SVM	Group suggestion [28]
		Explore context and content [52]
		Age and gender classification [23]
		Event relatedness classification [19]
	K-nearest-neighbor	Geo-location inference [31]
		Text classification [22]
	Naive Bayes	classify tweets reflecting students' problems [50]
		Classify Twitter user interests [51]
Unsupervised	K-means	Multimedia features clustering [29]
		Text clustering [22]
		Cluster a set of photo geographically [55]
		Clustering stability [56]
	Ensemble Clustering	Identify event on social media [20]
	Density-based clustering	Text-stream clustering for event [19]

2.3.1 Supervised Approach

The supervision in the learning approach comes from labeled class in the training data set [4]. Typically, data set in supervised approach is divided into two parts, training and testing data sets. The classification models are built from the training data and use the learned models for prediction. The test data is used for evaluation to obtain classification accuracy. Many supervised methods such as k-nearest neighbors, naive Bayes classification and support vector machines have been used in social user mining.

A k-nearest neighbors algorithm is a non-parametric method for classifying objects using training data and test data. The performance of this algorithm depends on two factors: selecting an appropriate value for the parameter k and a suitable similarity function. The k-nearest neighbors has been widely used for social user mining due to its simplicity, effectiveness and robustness. A k-nearest-neighbor method is used to estimate the geo-location of user from images [31], where a k-nearest-neighbor searching method is employed for visual matching and geo-inference. For each query image, an aggregate feature distance is used to find the nearest neighbors, and derives estimated geo-locations from those tagged nearest neighbors. It is important to select the right visual features for an effective matching of large visual contents. Furthermore, Yassine et al. [22] utilized k-nearest neighbor algorithm for text classification. The algorithm used to classify comments into three subjectivity levels: neutral, moderately subjective, and subjective.

Naive Bayes algorithm is tremendously appealing in many mining research because of its simplicity, elegance, and robustness. . It is widely used in areas such as text classification [49]. The author in [50] implemented a Naive Bayes multi-label

classification algorithm to classify tweets reflecting students' problems. Their study focused on engineering students' Twitter posts to understand issues and problems in their educational experiences. They found Naive Bayes classifier to be very effective on their data set compared with other state-of-the-art multi-label classifiers. Moreover, Naive Bayes algorithm use to classify Twitter user interests using time series generated from the contents of tweet streams [51]. They adopt a multinomial Naive Bayes model and their experiments shows that the series based classifiers outperform up to eight competing classification solutions significantly.

Support Vector Machine (SVM) is a popular machine learning method for classification and other learning tasks. SVM has been used to evaluate various social mining tasks [28, 52, 23]. One advantage of using SVM classifier is its superior performance. For that reason, Yu et al. [28] used SVM classifier to evaluate group suggestion using images. They built SVM classifier upon each visual feature to address the problem of having poor statistical modeling and extra computational cost. In addition, Gaussian Kernel is used in SVM by [52] to explore both the context and content information using a latent structure between the correlated semantic concepts for annotation.

Moreover, Support Vector Machine learning package, i.e., Liblinear [53], has been used by [23] for classification of Netlog posts according to age group and gender. They represented each Netlog post as a sparse binary vector for SVM classifier. Lee et al. [19] developed a supervised learning model to create extensible measure metrics for offline evaluation of event relatedness. Cheng et al. [13] used Weka toolkit [54] to

implement proposed classification algorithms that can identify word in tweets with a strong local geo-scope.

2.3.2 Unsupervised Approach

Unsupervised learning approach is designed to work with not labeled class data. In this approach, learning algorithms build a model based on the similarity or dissimilarity between data objects. Several unsupervised techniques have been evolving with social data mining.

K-means is a typical example of unsupervised learning method that attempts to find a user-specified number of clusters (k), which are represented as centroids. It has been broadly used on social user mining [55, 29, 22, 56]. Generally, the benefit of using this algorithm is its efficiency in processing large data sets, i.e., in many social user mining tasks. On the other hand, as social media data are noisy, we need to be very careful when we use this method. That is because it is sensitive to noise. In order to obtain an intuitive understanding of impact on ad click behavior, Cheng et al. [29] utilized various multimedia features. They quantize each multimedia feature into multiple bins using a k-means clustering algorithm. In addition, they use a Gaussian Mixture Component model to cluster images based on content similarity and use the cluster membership as the feature. One of the reasons that they use this model, instead of using more advanced models, is because of its scalability. In addition, k-means algorithm is adapted for training model of text subjectivity [22]. The goal of this study was to cluster text into different categories. Their experiments show high accuracy for the model in determining subjectivity of text as well as predicting friendship.

Moreover, k-means used for deriving meaningful data from unstructured text associated with geo-referenced data [55]. Ahern et al. [55] employs k-means clustering algorithm to cluster a set of photographs geographically. Geographical distance is used as the distance metric and the stopping condition for the k-means is reached when each cluster's centroid movement drops below 50 meters. In [56], clustering stability used the k-means clustering algorithm on the low-dimensional representation. To quantify how stable the clusters are, they use the standard precision and recall measures. Both precision and recall are constantly high across different data set and cluster sizes.

To identify event on social media, Becker et al. [20, 57] used ensemble clustering approach, which combines multiple partitions of a document [58]. They use weighted ensemble approach for clustering the social media documents, which collectively considers the rich features of the documents. The experiments show that the ensemble approach is effective to identify events. Furthermore, Lee [19] develops online text-stream clustering approach for event evaluation. His method includes three parts: a dynamic term weighting scheme, a sliding window model, and an online density-based clustering approach. One of the advantages of using density-based clustering approach is that it is based on density connectivity and treats noises as outliers which would not be involved in any cluster.

2.4 Discussion

In this section, we discuss about the current related work to social user mining, and right direction of it. First, we started by surveying the related work on the contents based social user mining. For this aspect, we summarizes the literatures by three data

types: textual, visual information, and both. Although many researchers have already studied different problems in social media mining using different data type, approaches that apply textual and visual data for social multimedia mining are still limited. Unlike the previous approaches that used one type of data, textual or visual, our main approach in this dissertation focused on utilizing the two data type together. Second, we reviewed the current work on the semantics based social user mining. Based on our survey, sometimes tags that annotated by users are ambiguous and overly personalize. To improve the quality of tagging, in this dissertation we use a semi-automatic tagging system that helps to tag a multimedia objects.

Lastly, we reviewed the social user mining based on data mining techniques. Different data mining techniques have been used in social media mining. We first survey the social user mining based on supervised approach, and then we investigate the social user mining based on unsupervised approach. Although researchers have already applied the data mining techniques to study different problems in social media mining, mining social media data is still very challenging due to the unstructured and noisy of social media data.

CHAPTER 3: MINING FRAMEWORK

3.1 Overview

It is important to develop an appropriate framework that is suitable for different data sources and types. Early work in social mining focused on developing a suitable framework that performed specific tasks of social user mining. The mining and analysis framework for evolving social data was proposed by [59].The framework designed to encapsulate steps needed to support social evolving data cleaning, summarizing and analysis without any data integration. In order to improve the quality and efficiency of the social user mining task, we propose a novel mining framework for the social media user, which includes the following components: the data assembly module for different media sources; data integration module; and mining applications.

3.2 Framework Structure

The proposed mining framework has three main modules: 1) data assembly module for different media source, 2) data integration module, and 3) mining applications. Figure 3.1 shows the overview of the mining framework. User generated data in social media is often highly unstructured. The main focus of the data assembly module is how to transform the collected unstructured data into structured data. For data integration, two methods were built: 1) content based data integration, and 2) semantic

based data integration. Lastly, two mining applications were utilized as an example of social user mining: 1) gender classification, and 2) user location prediction.

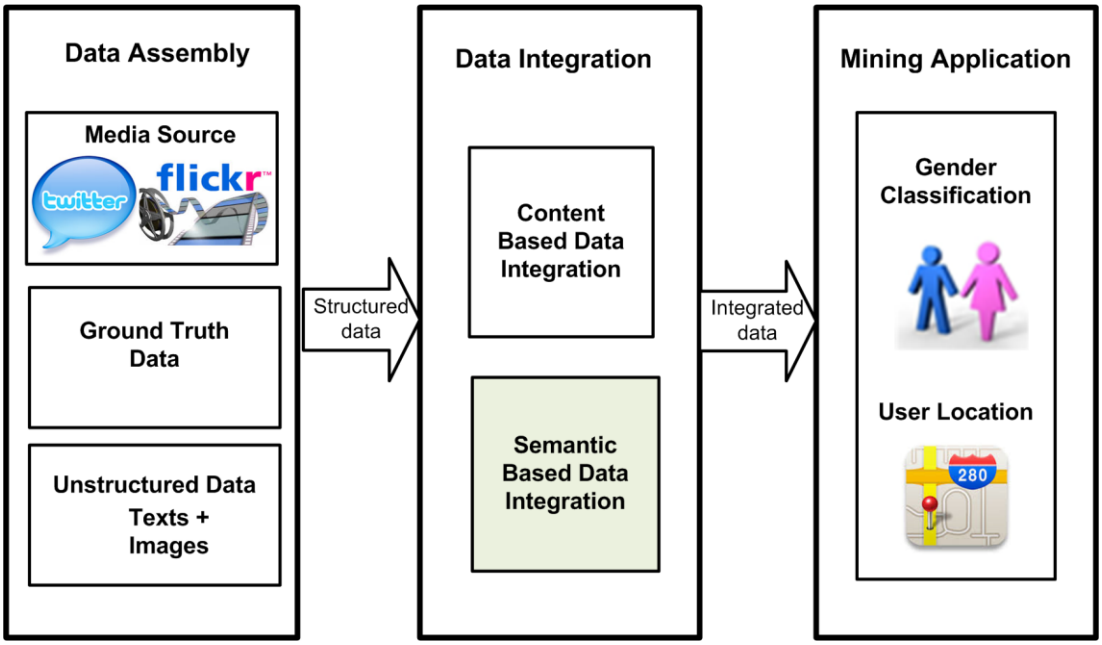

Figure 3.1. Overview of mining framework: data assemble, data integration, and mining application

3.3 Discussion

In the recent years, research has focused on acquiring a better understanding of social media data. Due to the nature of social media data, new methods and tools are required. Social user mining is an example of the new social media data category that researchers should developed. It is extremely important to integrate social media data with different data types because of the advancements and availability of various data in the social media networks. The proposed framework is designed to function with several different social media platforms, a variety of social media data, and various mining applications.

CHAPTER 4: DATA ASSEMBLE

4.1 Overview

In this chapter, we introduce a data assemble module that handles both textual and visual information from different media sources, and focuses more on evaluating the appropriate multimedia features for social user mining. The nature of social media data is significantly different from the data in traditional data mining because the social media data is big, noisy, highly unstructured, and incomplete compared to traditional data [8]. A major challenge exists in order to make proper structured data out of various media sources to discover unknown and meaningful information about users. In addition, the ability of users to adjust their privacy setting may produce missing information, which leads to another challenge. In many social media sources, it is hard to access the user information due to these privacy setting. Another challenging aspect of social media data is a lack of ground truth data. For any social user mining task, we are required to build a ground truth data set in order to evaluate the results. This task proves challenging because we need to have a label data in order to produce such a task. However, numerous users give incomplete or wrong information in their profile.

In order to overcome these challenges, we introduce a data module for social user mining to model both textual and visual data. Three main tasks are associated with this

module: (1) data collection, (2) data preprocessing, and (3) data representation. Depending on the data type, each task has a different way of processing data. Figure 4.1 illustrates the data assembly module.

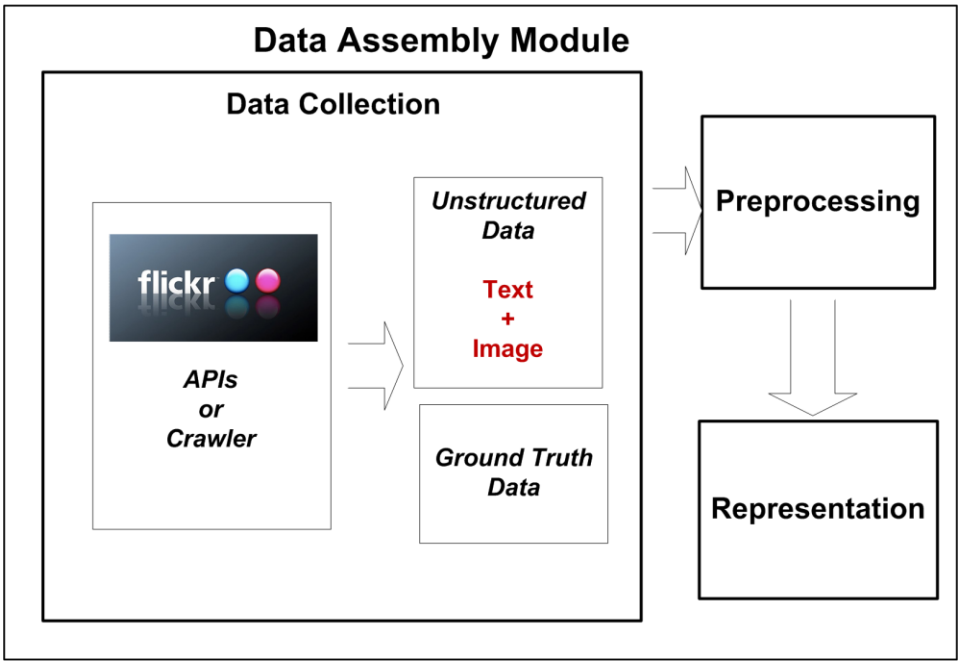

Figure 4.1. Data assembly module

4.2 Data Collection

The first task, in the data assembly module, is data collection. There are two ways to collect social user data: (1) by using crawler, and (2) by using APIs of each social media site. This section describes how we obtained Flickrs' data for our analysis. First, we studied the collection of the ground truth data by using crawler, and then followed with the collection of the textual and visual information by using Flickrs' API.

4.2.1 Ground Truth Data

In order to evaluate any experiment, we need to have ground truth data. Most of Flickr's registered users' list their profile information on a profile page. Each user's profile consists of some information set up by the user, such as demographic and geographic information. Figure 4.2 shows an example of Flickrs' user profile.

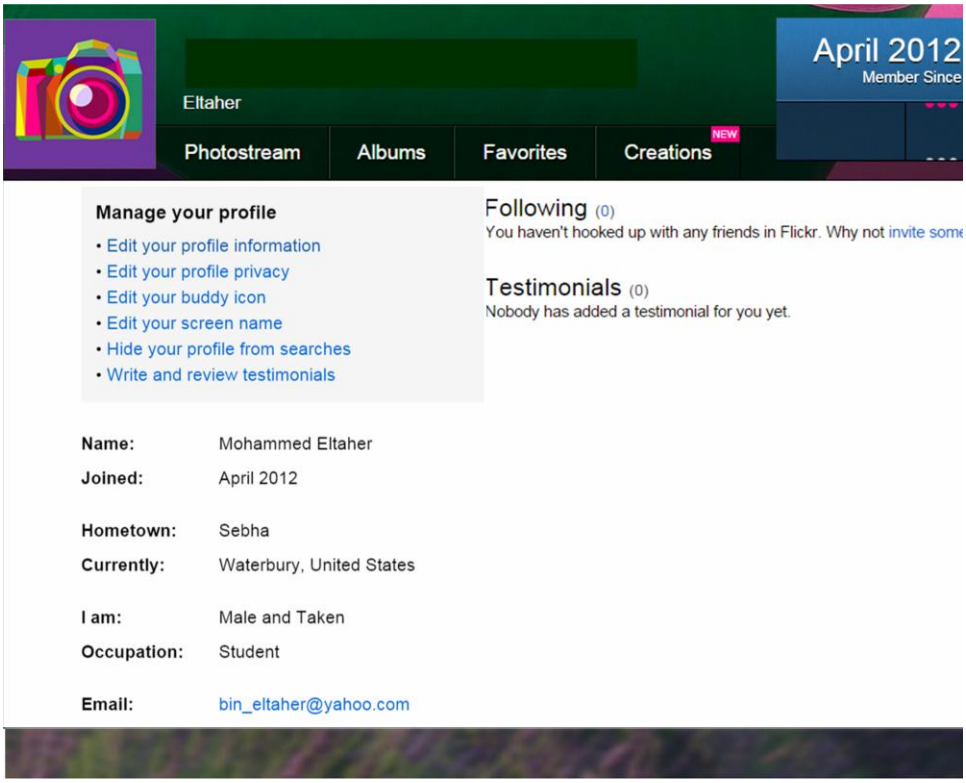

Figure 4.2 an example of Flickrs' user profile

We collected Flickr's users' profile information by using crawler as shown in Figure 4.3. Each of Flickr's user have a unique ID. Using these IDs, we first opened the user profile page. Then, we start to view the page source and extract the user profile attributes. In our research, we build a ground truth data based on 215k users with

different attributes including gender, location, marital status, country, city, hometown, and occupation. Table 4.1, below, shows the details of the ground truth data.

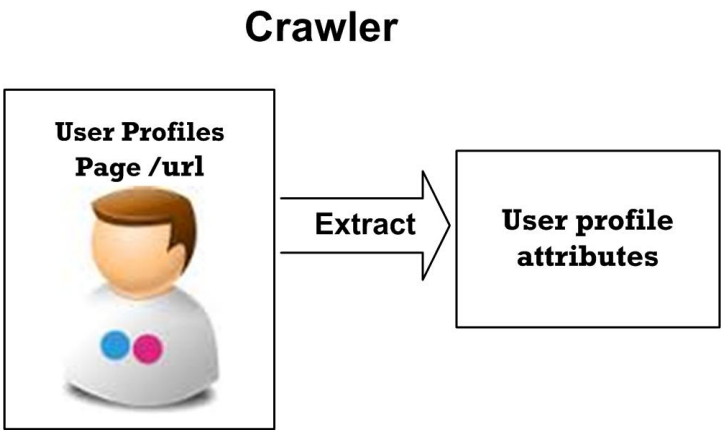

Figure 4.3 Crawler for user profile

Table 4.1 Ground truth data set

Attribute	*Number of users from 215k total*
Gender	148,511
Name	124,230
Marital status	98,475
Country	92,087
City	90,281
Hometown	82,221
Occupation	76,508

4.2.2 Textual and Visual Data

Flickr is one of the best online photo management and sharing sites on social media. In order to allow developers to access information, Flickr offers a comprehensive API that allows developers to create any application for their data. Textual and visual data can be obtained through Flickr's public API, which allows anybody to download information with the user's authorization.

The extracted data (images, textual annotations, and photographic meta-data) from Flickr can reveal information about users, including their interest, where they live and what they do. In our module, textual data was created by using tags from Flickr's users, and the visual data was represented through images. Figure 4.4 shows an example of Flickr's user's textual and visual data.

In addition, we use the content of Twitter's users' tweets as another example of textual information. For Twitter's data, we used Twitter data set collected by [13]. This data set was originally collected between September 2009 and January 2010 by crawling through Twitter's public timeline API as well as crawling by breadth-first search through social edges to crawl each user's followees/followers. The data set is supplementary divided into training and test sets. The training set consists of users whose location is set on city levels, within the US continental, resulting in 130,689 users with 4,124,960 tweets. The test set consists of 5,119 active users with approximately 1,000 tweets from each user. Here, each user's location was recorded as a coordinate (i.e., latitude and longitude) by a GPS device.

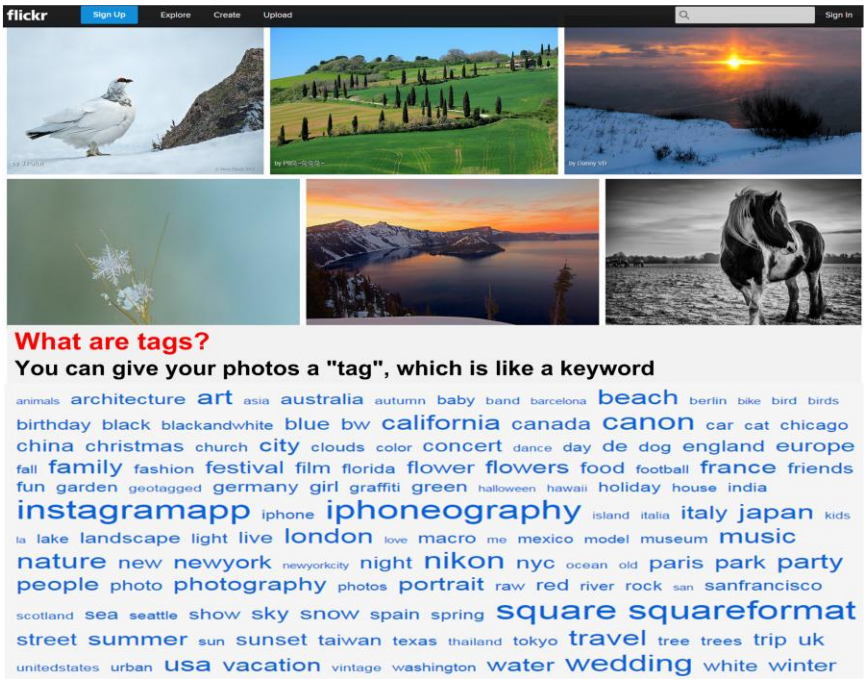

Figure 4.4. Textual and visual data example represented through user's tags and images of user

4.2.2 Semantic Data

In recent times, many social multimedia sharing networks, such as Flickr, have allowed their users to easily annotate multimedia objects with keywords. Tagging multimedia objects is defined as labeling images or videos with a set of keywords. However, sometimes, tags that are annotated by users, are ambiguous and overly personalized. A semi-automatic tagging system that helps to tag multimedia objects would improve the quality of tagging. In our research, in order to collect the semantic data, we used a semi-automatic image tagging system called *Akiwi*[4] to suggest keywords

[4] http://www.akiwi.eu/

for images. *Akiwi* uses an enormous collection of 15 million images tagged with keywords. Basically, *Akiwi* retrieves images that are visually very similar to the query image. Based on the keywords of these images, *Akiwi* tries to predict the keywords for the unknown image. Figure 4.5 shows an example of *Akiwi* with an untagged image, retrieved similar images, and suggested keywords.

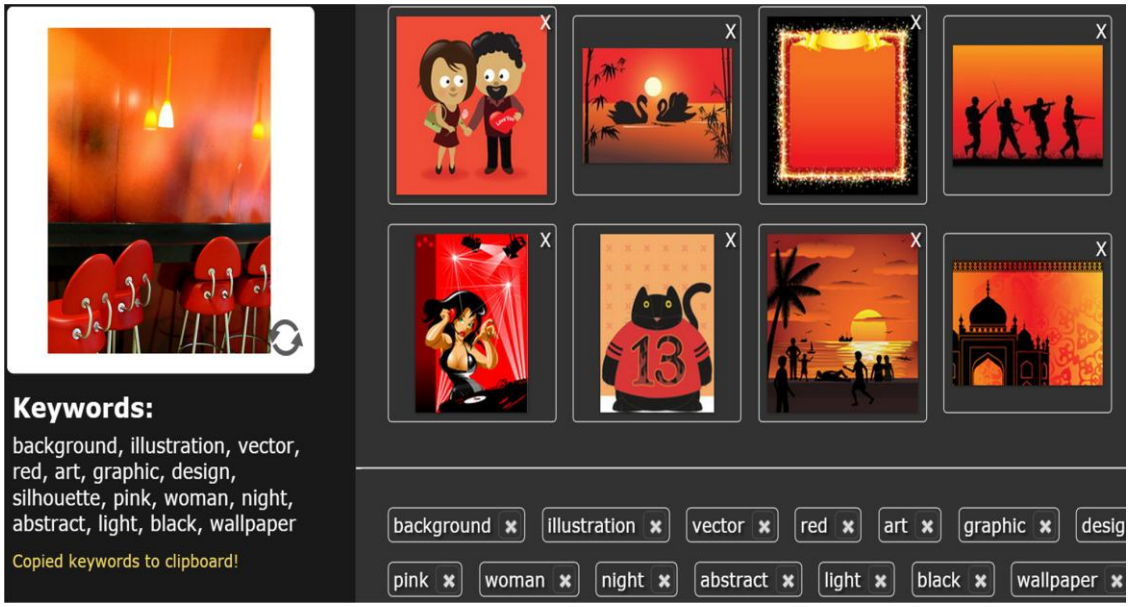

Figure 4.5. A semi-automatic image tagging system (Akiwi)

Akiwi is able to suggest keywords for uploaded untagged images. For each user, we have up to 50 images. One by one, we begin to query of all images per user to *Akiwi*. *Akiwi* retrieves images that are visually very similar to the query image. Based on the keywords of these images, *Akiwi* tries to predict the keywords for the unknown image. Figure 4.6, below, shows an example of the semantic data collection.

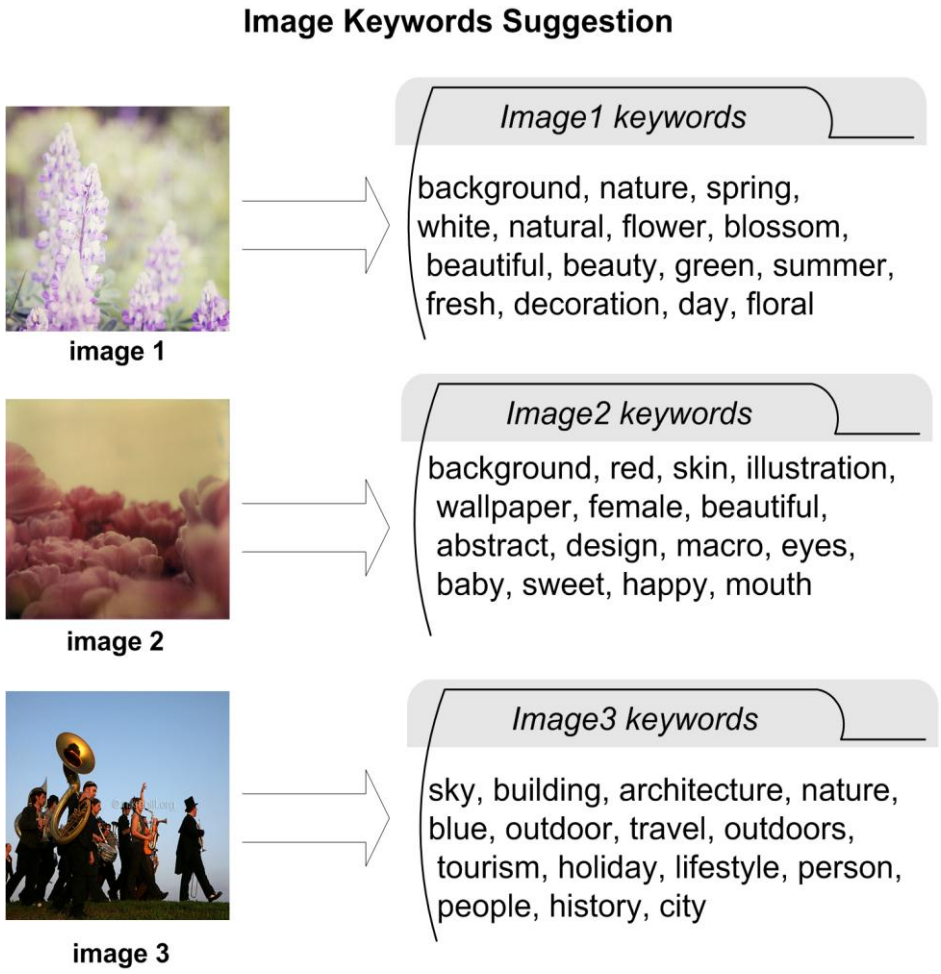

Figure 4.6 Example of semantic data collection

4.3 Preprocessing

The data preprocessing task aims to make the collected information more consistent in facilitating the representation of both textual and visual information. Depending on the data type and the application, the preprocessing method is different from one to others. For example, removing nonalphabetic characters (e.g., "@") and stop words is important for developing many application.

4.4 Representation

Multimedia data in social media is represented in many different ways. This section focuses on the representation of textual, visual, and semantic data.

4.4.1 Textual Data

The representation of textual information plays a critical role in many applications of social user mining. One of most commonly used and effective approaches of text representation is vector-space. The vector-space representation treats each document as a bag of words. In the bag of words approach, each document is represented as a set of words, and by the number of times each word occurs in the document. In other words, each word is represented as a separate variable having numeric features.

The text information used to describe the image by the users, such as tags, titles, descriptions and comments. In this study, we used tags as textual feature because it reflect what users consider important in their images, and also reveal the users' interest. This feature is denoted as $T = \langle t_1, t_2, \dots, t_n \rangle$

4.4.2 Visual data

Visual information on social media is usually represented by multiple features. A visual feature is defined to capture a certain visual property of an image, either globally for the entire image or locally for a small group of pixels. For example, an image can be represented by different features such as color, edges, and texture. The selection of appropriate features is beneficial for social user mining and hidden knowledge discovery.

Evaluating such appropriate visual features for social mining has not been sufficiently studied. Our main focus with the visual data is to evaluate different visual feature for social user mining based on images. Furthermore, we can also focus on examining what features we extract from images that allow us to explore more information about the user.

Color is one of the most widely used visual features and has been extensively studied on the multimedia mining research. Color descriptors can be used for representing the content of images. There are a number of color space are available such as, RGB, HSV, HSL and YUV. RGB stores individual values for red (R), green (G) and blue (B) for each pixel at (x,y). HSV (hue, saturation, value) and HSL (hue, saturation, lightness/luminance) are another color model that used in multimedia mining . In HSV, the brightness of a pure color is equal to the brightness of white whereas on HSL the lightness of a pure color is equal to the lightness of a medium gray. The YUV model labels a color space in terms of one luma (Y) and two chrominance (UV) components.

4.4.3 Semantic Data

Semantic information in social media is mainly represented based on images or videos. In order to represent the semantic data, we label the semantic content of user's images with set of keywords using a semi-automatic image tagging system. This feature is denoted as $K = \langle k_1, k_2, \dots, k_n \rangle$

4.4.4 Discussion

Overall, we show that social media sites contain substantial, useful information about user. With the data assemble module, developed in this dissertation, we can

effectively collect, preprocess, and represent different types of data across various social media sites. Regardless of the data types used or social media sites, this module can be helpful for many applications in social user mining.

CHAPTER 5: DATA INTEGRATION

5.1 Overview

By combining the features of image and textual attributes that are generated by user, interesting properties of social user mining are revealed. These properties serve as a powerful tool for discovering unknown information about the user. However, there is a minimum amount of research reported on the combination of images and texts for social user mining. The progress of data mining techniques makes it possible to integrate different data types in order to improve the mining tasks of social media, and thus make them more effective.

In social media networks, visual data, such as images, co-exist with text or other modalities of information. In order to benefit from different data modalities, further research is obligatory.

This chapter introduces the first step towards an integration module which is based on visual and textual data available in social media. We utilized tags and images of users with a novel approach of information fusion in order to enhance the social user mining. Here, two different approaches were applied to enhance social user mining: (1) contents based image fusion, and (2) semantic based image fusion.

5.2 Content Based Image Fusion

Through the content based image fusion, we combined textual and visual information by using image contents. We proposed a data integration method between the user's tags and image contents. For the image contents, we used a hue histogram and a hue in bag of words. We implemented the tags with hue histogram as a feature vector as well as tags with hue in bag of words.

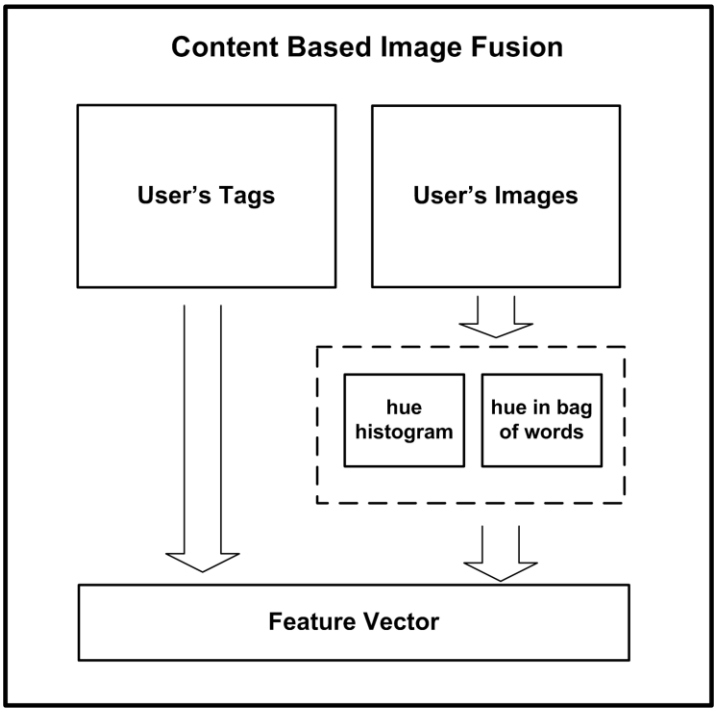

Figure 5.1. Content based data integration module

5.2.1 Integrated Data Units

Flickr allows their users to annotate their photos with textual labels called "tags". Tags, in social media, are accurate descriptors of content, and could be used in many

mining applications [60]. In this section, we captured the content of users' Flickr photos through the user-generated tags. The first element of our content based data integration module is the user's tags. For each user u, we utilized up to 300 tags whereby each tag represented as a word denoted below:

$T_u = \langle t_1, t_2, \dots, t_n \rangle$, where n is the number of tags for each user .

For the second element of the proposed content based data integration module, we use the user's images as visual data. Specifically, we represent the images through two features known as hue histogram and hue in bag of words. Hue histogram is based on the hue value of all the pixels in the user's images. Each hue value in the HSL or the HSV color space represents an individual color. For the hue in bag of words feature, we selected the top two colors by assigning "1" to the feature value for these top colors and "0" to the others. The hue histogram and hue in bag of words can be denoted as below:

$HS_u = \langle hs_1, hs_2, \dots, hs_n \rangle$, where n is the number of colors for each user u.

$HBW_u = \langle hbw_1, hbw_2, \dots, hbw_n \rangle$, where n is number of color for each user u.

5.2.2 Integration Scheme

We continued to implement the users' tags with the hue histogram and the hue in bag of words as a feature vector. Figure 5.2 shows the scheme of the content based image fusion. For the tag features, each user u has a feature vector F_t that corresponds to all the users' tags. This feature vector can be defined as:

$F_t = <T_u>$, where T_u is the users' tags.

For the hue histogram, each user has up to 50 image. We calculated the hue histogram for each user based on their images, and determined the average based on 50 images for each color. For the hue in bag of words, we selected the top two colors for each user based on the images. The feature vectors of the hue histogram and the hue in bags of words can be defined as:

F_{hs}= <HS_u>, where HS_u is hue histogram per user.

F_{hbw}= <HBW_u>, where HBW_u is a hue in bag of words per user.

Tag Features	*Hue histogram Features*	*Hue in bag of words features*
$\downarrow F_t$	$\downarrow F_{hs}$	$\downarrow F_{hbw}$
$T_u = \langle t_1, t_2, \dots, t_n \rangle$	$HS_u = \langle hs_1, hs_2, \dots, hs_n \rangle$	$HBW_u = \langle hbw_1, hbw_2, \dots, hbw_n \rangle$

Figure 5.2. Feature vector of the content based data integration

5.3 Semantic Based Image Fusion

In this section, we studied the problem of integrating textual and visual data to perform social user mining tasks. We determined that the integration of the two data types will be more beneficial than using an individual data type. We proposed a data integration module that combined both textual and visual information. First, we applied a semi-automatic image tagging system called *Akiwi* to suggest keywords for images. *Akiwi* uses an enormous collection of 15 million images tagged with keywords. Basically,

Akiwi retrieves images that are visually very similar to the query image. Based on the keywords of these images, *Akiwi* tries to predict the keywords for the unknown image. Then, we integrated these keywords for individual users' tag. Figure 5.3 illustrated the data integration module.

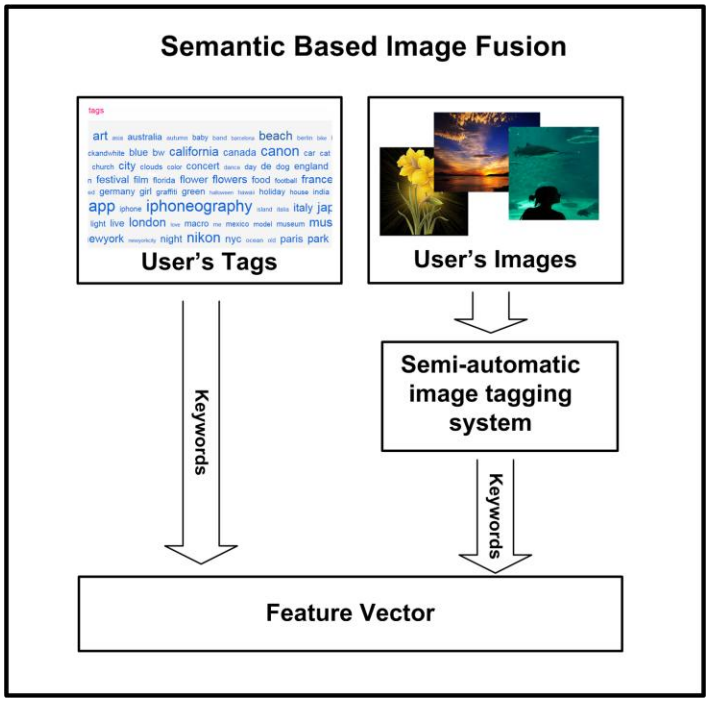

Figure 5.3 Semantic based data integration module

5.3.1 Integrated Data Units

Social media networks provide their users with the ability to describe any photos' contents by manually annotating the photos. This process is called "tagging". Similar to the above represented content based integration module, our first element of the semantic based data integration module is users' tags. Some users' tags are unreliable due to excess noise in tags provided by users. These tags prove to be irrelevant or incorrectly spelled.

For example, only about 50% of the tags provided by Flickrs' users are in fact related to the images [61]. Due to tagging inaccuracies, we use a semi-automatic image tagging system to suggest keywords for our images. These keywords are considered as the second element of our proposed semantic based data integration module. We applied *Akiwi* to suggest keywords for the images. *Akiwi* uses an enormous collection of 15 million images tagged with keywords. Basically, *Akiwi* retrieves images that are visually very similar to the query image. Based on the keywords of these images, *Akiwi* tries to predict the keywords for the unknown image.

5.3.2 Integration scheme

For the semantic based data integration module, we implemented the users' tags with the keywords retrieved from *Akiwi* as a feature vector. The main difference in this module focuses on tags and keywords generated by users, as opposed to keywords generated by *Akiwi*. Figure 5.4 show the scheme of the semantic based image fusion. This feature vectors of tags and keywords for each user can be defined as:

F_t= <T_u>, where T_u is users' tags.

F_k= <K_u>, where K_u is users' keywords.

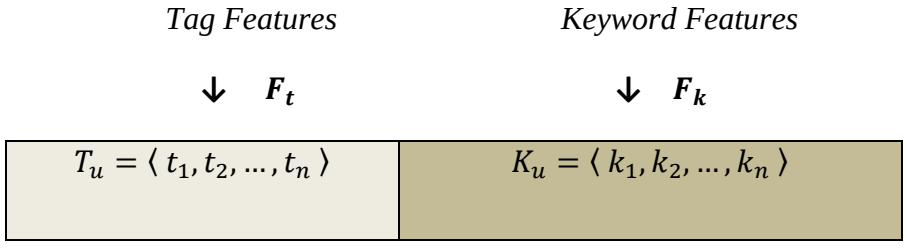

Figure 5.4 Feature vector of the content based data integration

5.4 Conclusion

In this chapter, an integration module, based on visual and textual social media data were presented. We utilized tags and images of users with a novel approach of information fusion in order to enhance the social user mining. We proposed two different approaches: (1) contents based image fusion and (2) semantic based image fusion. In the next chapter, we utilize these two approaches with a mining application to demonstrate how our new semantic based approach outperforms the content based approach.

CHAPTER 6: MINING APPLICATION

In this chapter, we introduce two different examples of social user mining applications: gender classification and user location.

6.1 Gender Classification

6.1.1 Overview

The gender classification problem in Flickr is one of the applications that our framework addresses. The authors in [62] introduced a gender identification technique for Flickr's users based only on tags. As compared to our approach, where we apply tags and images. For a user u, given his d_u (tags and images) from Flickr, we predict the gender of u based on tags, images, and a combination of both tags and images.

6.1.2 Classification Algorithms

Generally, there are various mining techniques that can be used in evolving social user mining. To solve the gender classification problem, we selected two popular classifiers: the Naive Bayes and the Support Vector Machine.

The Naive Bayes classifier which is one of the most efficient and effective inductive learning algorithms for machine learning and data mining [63]. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem

with the "naive" assumption of independence between every pair of features. In this experiment, we adopted a multinomial Naive Bayes model. This model implements the Naive Bayes algorithm for multinomial distributed data, where the data is typically represented as vector. Given the gender classification problem having G classes $\{g_1, g_2\}$ with probabilities $P(g_1)$,$P(g_2)$, we assign the class label G to a Flickr user u based on the feature vector $D_u = (d_1, d_2, \dots, d_N)$, where d_N represent the user data such as tags and images:

$$g = arg \max_{g} P(g|D_u)$$

The above equation is to assign a class with the maximum probability given the feature vector of user data D_u. This probability can be formulated by using Bayes theorem as follows:

$$P(g|D_u) = \frac{P(g) \times \prod_{i=1}^{N} P(d_i|g)}{P(D_u)}$$

Here, the objective is to predict the most possible class to the user u giving the feature vector D_u that contains N features to the most possible class.

The Support Vector Machine (SVM) is a popular machine learning method for classification and other learning tasks [64]. In our experiment, we adopted the C-Support Vector Classification (SVC) which is implemented based on libsvm [65]. The main idea of applying SVM on classification is to find the maximum-margin hyperplane to separate among classes in the feature vector space. Given a set of Flickr data D_u that is relevant to a user u and class labels for training$\{(d_u, g_u)|u = 1, \dots, N\}$, where d_u is feature vectors

of user data and g_u is the target class label, SVM will map these feature vectors into a high dimensional space.

6.1.3 Content Based Classification

As the amount of social media contents grows, researcher should identify robust ways to discover unknown knowledge about users, based on these contents. In this section, we explain how the content of tags and images are used for gender classification.

6.1.3.1 Tag Data

Tags reflect what users consider important in their images, and also reveal the users' interest. We assume that male and female tagging vocabularies are different, and this difference can be used to identify their gender. To test our assumption, we built a dictionary containing female and male tagging vocabularies. In order to determine the importance of tags, we compute the gender tagging vocabulary by counting the number of different gender users who used the respective tag. Then, we calculated the probability of a gender given the utilized tags. Table 6.1 shows a sample of gender probability over the tags.

6.1.3.2 Image Data

The color histogram is a representation of color distribution in an image. For image data, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges that span the color space for the image. The color histogram can be built for any kind of color space. The hue histogram is based on the hue value of

all the pixels in image. Each Hue value in HSL or HSV color space represents a color by itself. Table 6.2 represents our example of the hue histogram for a 32 pixel image.

Table 6.1: Tag gender dictionary

Tag	*Male frequency*	*P(male | tag)*	*Female frequency*	*P(female | tag)*
panorama	6921	0.785	1896	0.215
cupcakes	776	0.309	1738	0.691
lake	9887	0.628	5869	0.372
fisherman	2125	0.67	1045	0.33
piazza	1085	0.679	514	0.321
dessert	1815	0.442	2290	0.558
soft	2012	0.452	2436	0.548
police	4350	0.728	1623	0.272
sisters	1108	0.408	1609	0.592

Male frequency: number of male users that have used the tag at least once
Female frequency: number of feamle users that have used the tag at least once

Table 6.2 Example of Hue histogram

	Value						
Histogram	10	5	2	5	7	1	2
Color	Red	Orange	Yellow	Green	Blue	Indigo	Violet
Feature	0.3125	0.1562	0.0625	0.1562	0.2187	0.0312	0.0625

The hue in bag of words approach is motivated by an analogy of learning methods that applies the bag-of-words representation for text categorization [66], visual categorization with bags of keypoints [67], and bags of features [68]. In this approach, we selected the top two colors by assigning "1" to the feature value for these top colors and "0" to the other colors. Considering the above example with hue histogram, Table 6.3 shows how we used the hue in bag of word. In this case, the top two colors are red and blue.

Table 6.3 Example of Hue in bag of words

	Value						
Histogram	10	5	2	5	7	1	2
Color	Red	Orange	Yellow	Green	Blue	Indigo	Violet
Feature	1	0	0	0	1	0	0

6.1.4 Semantic Based Classification

Social user mining research better understands the semantic content of multimedia data added by the user. However, manually annotating images requires time and effort, and it is difficult for users to provide all relevant tags for each image. Thus, a semi-automatic image tagging system emerged and has recently involved. To improve the quality of tags, we applied a semi-automatic image tagging system called *Akiwi* to suggest keywords for images .The goal of using semi-automatic image tagging system is to assign a few relevant keywords to the image to reflect its semantic content. This

process improves the quality of tags by utilizing image content. For the gender classification problem, the semantic based approach is conducted based on the collected keywords from *Akiwi*.

6.1.4.1 Keyword Data

Similar to the tags data in the content based classification, the assumption is that male and female keyword vocabularies are different. This difference can be used to classify their gender. To test our assumption, we built a dictionary containing female and male keyword vocabularies. In order to determine the importance of keywords, we compute the gender tagging vocabulary by counting the number of different gender users who used the respective keywords. Then, we calculated the probability of a gender given the utilized keywords.

6.1.4.2 Keyword and Contents Based Data

As we introduced in chapter 5, we proposed a data integration module to combine both semantic based and content based data. Particularly, we utilize this module for the gender classification problem by combined the keywords of the user with his/her tags. We use a feature vector to merge both the keywords and tags of the user.

6.1.5 Experiments & Discussion

This experiment utilizes Scikit-Learing tools in Python [69]. Two different classification methods, i.e., Naive Bayes and Support Vector Machine were used. In this experiment, we adopted the multinomial Naive Bayes model. This model implements the

Naive Bayes algorithm for multinomial distributed data, where the data are typically represented as feature vector. For the SVM, we adopted C-Support Vector Classification SVC which is implemented based on libsvm. For both classifiers, we use the fit (X, Y) method. This method fit the classifier according to the given training data. Next, we used the predict(X) method to perform the classification in a sample of X. In our case, X represents the feature matrix of the data, while Y represents the user label.

6.1.5.1 Content Based Experiments

For the content based experiment, we implemented a multinomial Naive Bayes classifier. We examined the performance of different features, and difference appeared in the classification. Table 6.4 below shows the result of different features such as tags, hue histogram, and hue in bag of words. To assess the performance of our model, we used the standard classification accuracy (*Acc*) and F1 score as defined in the equations 6.1 and 6.2, shown below. For evaluation purposes, all classes are grouped into four categories: 1) true positives (TP), 2) true negatives (TN), 3) false positives (FP), and 4) false negatives (FN). For instance, the true positives are the users that belong to the positive class and are in fact classified to the positive class. Whereas the false positives are the users not belonging to the positive class but incorrectly classified to the positive class.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad (6.1)$$

$$F1 = \frac{2TP}{(2TP + FP + FN)} \qquad (6.2)$$

Table 6.4 Experiment results for content based classification

Features	***Accuracy***	***F1***
tags	0.7362	0.7349
huehist	0.6141	0.6140
huebow	0.5866	0.5786
tags+huebow	0.7365	0.7351
tags+huehist	0.7251	0.7228
huehist+huebow	0.6151	0.6150
tags+huehist+huebow	0.7181	0.7141

huehist: hue histogram
huebow: hue in bag of words

6.1.5.2 Semantic Based Experiments

To compare the performance of our approach, we use the classification accuracy (*Acc*), precision (*Pre*), and recall (*Rec*) metrics as well as F1 score as defined in the following equations:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \qquad (6.3)$$

$$\Pr e = \frac{TP}{TP + FP} \qquad (6.4)$$

$$\mathrm{Re}c = \frac{TP}{TP + FN} \tag{6.5}$$

$$F1 = 2\left(\frac{\mathrm{Pr}e \times \mathrm{Re}c}{\mathrm{Pr}e + \mathrm{Re}c}\right) \tag{6.6}$$

Where *TP* is true positives, *TN* is true negatives, *FP* is false positives, and *FN* is false negatives.

We performed the experiments by sampling of the data set for different features and classifiers, and then tested the performance of each classifier and each feature. The results are presented in Table 6.5. As seen in the table, the results show over 80% in terms of accuracy for gender classification when using keywords with both classifiers. This indicates that the proposed semantic based approach outperforms the content based one. In term of classifier, we observed that Naive Bayes is slightly better than SVM, specifically with tags. This is because the Naive Bayes classifier can work better even if there are some missing data.

Table 6.5 Experiment result for semantic based classification

Features	**Approach**	**Acc**	**Pre**	**Rec**	**F1**
Keywords	NB	0.82	0.81	0.82	0.81
	SVM	0.82	0.83	0.82	0.80
Tags	NB	0.78	0.82	0.78	0.78
	SVM	0.74	0.55	0.74	0.63
Keywords+ Tags	NB	0.80	0.80	0.80	0.79
	SVM	0.78	0.61	0.78	0.68

6.2 User Location

6.2.1 Overview

Knowing users' home locations in social media networks is an importance for many applications such as location-based marketing and personalization. Although many of social media networks allow their users to specify their locations along with other demographics information, still many users do not provide such information because of privacy concern or others. Therefore, it is an important task to be able to automatically discover users' home locations using their social media data. In this section, we focus on the case of Twitter users and try to predict their city locations based on only the contents of their tweet messages. For a user$u$, given a set of his/her tweet messages $Tu = \{t_1, \ldots, t_{|T_u|}\}$, where t_i is a tweet message up to 140 characters, and a list of candidate cities, C, predict a city c that is most likely to be the home location of u.

6.2.2 Gaussian Mixture Model (GMM)

GMM is a mature and widely used technique for clustering, classification, and density estimation. It is a probability density function of a weighted sum of a number of Gaussian components. This model assumes all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. To address the location problem, we propose to use the bivariate Gaussian Mixture Model (GMM) as an alternative to model the spatial word usage and to estimate$P(C|w)$.

6.2.3 Modeling Location

In the recently years, many generative methods have been proposed to solve the proposed problem such as in [70, 71, 72]. Assuming that each tweet and each word in a tweet is generated independently, the prediction of home city of user u given his or her tweet messages is made by the conditional probability under Bayes rule and further approximated by ignoring $P(T_u)$ that does not affect the final ranking as follows:

$$P(C|T_u) = \frac{P(T_u|C)P(C)}{P(T_u)}$$

$$\alpha\ P(C) \prod_{t_j \in T_u} \prod_{w_i \in t_j} P(w_i|C)$$

Where w_i is a word in a tweet t_j. If $P(C)$ is estimated with the maximum likelihood, the cities that have a high usage of tweets are likely to be favored. Assuming a uniform $P(C)$, we propose another approach by applying Bayes rule to the $P(w_i|C)$ of above formula and replace the products of probabilities by the sums of log probabilities:

$$P(C|T_u)\ \alpha\ P(C) \prod_{t_j \in T_u} \prod_{w_i \in t_j} \frac{P(C|w_i)\ P(w_i)}{P(C)}$$

$$\propto \sum_{t_j \in T_u} \sum_{w_i \in t_j} \log P(C|w_i)\ P(w_i)$$

Accordingly, give C and T_u, the home location of the user u is the city c ($\in C$) that maximizes the above equation as:

$$argmax_{c \in C} \sum_{t_j \in T_u} \sum_{w_i \in t_j} \log P(c|w_i)\, P(w_i)$$

We refer this model as spatial word usage. It suggest to estimate the city distribution on the user of each word, $P(C|w_i)$, and aggregate all evidence to make the final prediction. This model is similar to the one used in [13], where the focus was on the observation rather than derived theoretically. One of the common way to estimate $P(C|w)$ and $P(w|C)$ is the Maximum Likelihood Estimation (MLE). Yet, it suffers from data sparseness problem that underestimates the probabilities of words of low or zero frequency. Different smoothing techniques are proposed such as Dirichlet and Absolute Discount [73]. For better estimation, we use Gaussian Mixture Model (GMM) to improve the prediction while addressing the sparseness problem.

We use the bivariate Gaussian Mixture Model (GMM) as an alternative to model the spatial word usage and to estimate $P(C|w)$. formally, using GMM, the probability of a city c on tweeting a word w is as following:

$$P(c|w) = \sum_{i=1}^{K} \pi_i \, N(c|\mu_i, \sum_i)$$

Where each $N(c|\mu_i, \sum_i)$ is a bivariate Gaussian distribution with the density as:

$$\frac{1}{2\pi|\sum_i|^{1/2}} \, exp\left\{-\frac{1}{2}\,(c-\mu_i)^T \sum\nolimits_i^{-1} (c-\mu_i)\right\}$$

Where K is the number of components and $\sum_{i=1}^{K} \pi_i$ = 1. In order to estimate $P(C|w)$ with GMM, each occurrence of the word w is seen as a data point(log, lat), the coordinate of the location where the word is tweeted. For instance, if a user has tweeted

$phillies$ 3 times, there are 3 data points (i.e., (log, lat)) of the user location in the data set to be estimated by GMM.

6.2.4 Local Word Selection

We propose unsupervised measures to evaluate the usefulness of tweet words for location prediction task. We assumed that words have some locations of interests where users tend to tweet extensively. However, not all words have a strong pattern. For example, if a user tweets $phillies$ and $libertybell$ frequently, the probability for Philadelphia to be her home location is likely to be high. On the other hand, even if a user tweets words like restaurant or downtown often, it is hard to associate her with a specific location. That is because such words are commonly used and their usage will not be restricted locally. Therefore, excluding such globally occurring words would likely to improve overall performance of the task.

For local words selection, [13] used a supervised classification method. They manually labeled around 19,178 words in a dictionary as either local or non-local and used the frequency of a word as features to build a supervised classifier. Regardless of the promising results, such a supervised selection approach is challenging not only because of their labeling process to manually create a ground truth is labor intensive, it is also hard to transfer labeled words to new domain or data set. In addition, the dictionary used in labeling process might not differentiate the evidences on different forms of a word. As a result, a better approach is to automate the process such that the decision on the localness of a word is made only by their actual spatial word usage, rather than their semantic meaning being interpreted by human labelers.

Toward this challenge, we propose two unsupervised methods to select a set of "local words" from a corpus using the evidences from tweets and their tweeter locations directly. In the first method, the aim is to find the local words by non-localness. This is an automatic way to filter noisy non-local words out from given corpus. We use the stop words as counter examples. The local words tend to have the farthest distance in spatial word usage pattern to stop words. We first estimate a spatial word usage $p(C|w)$ for each word as well as stop words. The similarity of two words, w_i and w_j , can be measured by the distance between two probability distributions, $p(C|w_i)$and $p\left(C\middle|w_j\right)$. In the second method, we use a geometric-localness to find the local word. Basically, if a word w has a smaller number of cities with high probability scores, and smaller average inter-city geometric distances among those cities with high probability scores, then we can view w as a local word.

6.2.5 Experiments & Discussion

For validating the proposed ideas, we used the Twitter data set collected and used by [13]. This data set was originally collected between Sep. 2009 and Jan. 2010 by crawling through Twitter's public timeline API as well as crawling by breadth-first search through social edges to crawl each user's followees/followers. The data set is supplementary divide into training and test sets. The training set consists of users whose location is set in city levels and within the US continental, resulting in 130,689 users with 4,124,960 tweets. The test set consists of 5,119 active users with around 1,000 tweets from each, whose location is recorded as a coordinate (i.e., latitude and longitude) by GPS device, a much more trustworthy data than user-edited location information.

In this experiments, we considered only 5,913 US cities with more than 5,000 of population in Census 2000 U.S. Gazetteer. We preprocess the training set by removing nonalphabetic characters (e.g., "@") and stop words, and selects the words of at least 50 occurrences, resulting in 26,998 unique terms at the end in our dictionary. For the evaluation, we use two metrics a defined below. First, the accuracy (ACC) measures the average fraction of successful estimations for the given user set U. The successful estimation is defined as when the distance of estimated and ground-truth locations is less than a threshold distance $d = 100\,(miles)$. Second, for understanding the overall margins of errors, we use the average error distance *(AED).*

$$ACC = \frac{|\{u|u \in U \text{ and } dist(Loc_{true}(u), Loc_{est}(u)) \leq d\}|}{|U|}$$

$$AED = \frac{\sum_{u \in U} dist(Loc_{true}(u), Loc_{est}(u))}{|U|}$$

Table 6.6 shows the top-30 local words with GMM. This is when resorted by frequency from 3,000 NL-selected words. Note that most of these words are *toponyms*, i.e., names of geographic locations, such as *nyc*, *dallas*, and *fl*. Others include the names of people, organizations or events that show a strong local pattern with frequent usage, such as *obama*, *fashion*, or *bears*. Therefore, it appears that *toponyms* are important in predicting the locations of Tweeter users. Interestingly, a previous study in [70] showed that toponyms from image tags were helpful, though *not* significantly, in predicting the location of the images.

Table 6.6. The top-30 frequency resorted local words (GMM, NL).

la	nyc	hiring	dallas	francisco
obama	fashion	atlanta	houston	denver
san	diego	sf	austin	est
chicago	los	seattle	hollywood	yankees
york	boston	washington	angeles	bears
ny	miami	dc	fl	orlando

The baseline results are for our experiments are shown in table 6.7. This table presents the results of different models for location estimation. All the probabilities are estimated with MLE using all words in our dictionary. The baseline Models (1) and (2) (proposed by [13]) utilize the spatial word usage idea, and have around 0.1 of ACC and around 1,700 miles in AED. The Model (3) is a language model approach. This model shows a much improved result about two times higher ACC and AED with 400 miles less.

Table 6.7. The baseline Experiments Result

	Probabilistic Model	*ACC*	*AED*
(1) Proposed	$\sum\sum \log(\ P(c\|w_i)\ P(w_i))$	0.1045	1,760.4
(2) Chen et al.	$\sum\sum P(c\|w_i)\ P(w_i)$	0.1022	1,768.73
(3) Language Model	$\sum\sum \log P(w_i\|c)$	0.1914	1,321.42

Table 6.8 below show the overall result based on two different models, GMM and MLE with the best setting for each model. In terms of selecting local words, NL works better than GL in general.

Table 6.8 Location Estimation Result

Model	*Measure*	*#Word*	*ACC*	*AED*
GMM	NL	2000	0.486	583.2
GMM	NL	50*	0.446	509.3
MLE	GL	2000	.449	611.6

(First 50 words from top 2000 local words resorted by frequency)*

Table 6.9 below illustrates examples where cities are predicted effectively by using GMM estimation and NL-selected local words. Note that words such as *audition* (i.e., the Hollywood area is known for movie industries) and *kobe* (i.e., name of the basketball player based in the area) are a good indicator of the city of the Twitter user.

Table 6.9 Example of correctly estimated cities and corresponding tweet massages (local words are bold face)

Est. City	*Tweet Message*
Los Angeles	i should be working on my monologue for my **audition** Thursday but the thought of memorizing something right now is crazy
Los Angeles	i knew deep down inside ur powell s biggest fan p **lakers** will win again without **kobe** tonight haha if morisson leaves **lakers** that means elvan will not be rooting for **lakers** anymore
New York	the march **vogue** has caroline trentini in some awesome **givenchy** bangles i found a similar look for less an intern from teen **vogue** fashion dept just e mailed me asking if i needed an assistant aaadorable

CHAPTER 7: CONCLUSIONS AND FUTURE WORK

7.1 Research Contributions

Our dissertation highlights the need for Social user mining in view of the rapidly growing amounts of social network data. We have pointed out the unique characteristics of different social data that bring with new challenging and interesting research issues to be resolved.

A novel mining method for social users mining using different social media data was presented. This method contains a data assemble module for different media source, a data integration module, and mining applications. First, we introduced a data assemble module to handle both textual and visual information from different media sources, and then focused on evaluating the appropriate multimedia features for social user mining. By using the data module, we can effectively collect, preprocess, and represent different types of data across various social media networks.

A new data integration method was proposed to integrate textual and visual data. Unlike the previous approaches that used a content based approach to merge multiple types of features, our main approach is based on image semantic through a semi-automatic image tagging system. Our methods were applied to two mining applications: user location and gender classification, and showed the performance of our methods to discover unknown knowledge about user. For gender classification, we performed the experiments with the data set, and the results indicate over 80% in terms of accuracy for

gender classification, which outperforms the content based approach.

7.2 Publications

A list of related publications with respect to the algorithms and methodologies as mentioned in this dissertation can be found as following:

7.2.1 Journal

1. Mohammed Eltaher and Jeongkyu Lee. "Social User Mining: Survey on Mining Different Types of Social Media Data," *International Journal of Multimedia Data Engineering and Management (IJMDEM)* 4 (2013).

7.2.2 Conference

1. Hau-Wen Chang, Dongwon Lee, Mohammed Eltaher, Jeongkyu Lee"@phillies tweeting from philly? predicting twitter user locations with spatial word usage," *in Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, Aug 2012.
2. Mohammed Eltaher, Jeongkyu Lee" User Profiling of Flickr: integrating multiple types of features for gender classification," *International Conference on Software and Information Systems (ICSIS), Las Vegas,USA* , May 2015. Accepted
3. Mohammed Eltaher, Yan Chen, Yawei and Jeongkyu Lee" Gait-Based Gender Classification using Kinect Sensor," *ASEE Annual Conference and Exposition, Seattle ,USA* , June 2015. Accepted

4. Mohammed Eltaher, *Haiyang Wang* and Jeongkyu Lee. “Motion Detection Using Kinect Device for Controlling Robotic Arm,” Proceeding of 2013 New England American Society for Engineering Education Conference, Northfield, VT, April 27-28, 2013

7.2.3 Relevant Poster Award

1. Graduate student poster competition (incredible mention) Mohammed Eltaher and Jeongkyu Lee. Social User Mning, 2014 UB Faculty Research Day, Bridgeport, CT, March 28, 2014

7.3 Future work

Many directions are worth further pursuing in future. For the social user mining, we believe there are various ways to further improve to discover more information about user using social media data. For example, for the mining applications that we investigated in this dissertation, we focused just in two application, gender classification and user's location. We can further investigate another mining application such as age prediction and others. Moreover, by utilize the social media data, there will always be new opportunities to discover unknown knowledge about user in social media.

Another direction to improve the social user mining research is to combine more multiple type of data for any mining application. In this dissertation, we only consider tags and images of Flickr's users as well as tweet content of Twitter's users for different

predication tasks. By integrating multiple type of social media data, we can address many other challenges and opportunities to improve the social user mining research.

REFERENCES

[1] M. Naaman, "Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications," Multimedia Tools Appl., vol. 56, no. 1, Jan. 2012, pp. 9–34.

[2] C. C. Miller, "Why twitter's ceo demoted himself," New York Times, October, vol. 30, 2010.

[3] S. Boll, "Multitube–where web 2.0 and multimedia could meet," MultiMedia, IEEE, vol. 14, no. 1, Jan 2007, pp. 9–13.

[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[5] P.-N. Tan, M. Steinbach, V. Kumar et al., Introduction to data mining. Pearson Addison Wesley Boston, vol. 1, 2006.

[6] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip et al., "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, 2008, pp. 1–37.

[7] M.-F. Moens, J. Li, and T.-S. Chua, Mining User Generated Content. Chapman & Hall/CRC, 2014.

[8] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: A survey," SIGKDD Explor. Newsl., vol. 15, no. 2, Jun. 2014, pp. 20–29.

[9] R. Zafarani, M. A. Abbasi, and H. Liu, Social media mining: an introduction. Cambridge University Press, 2014.

[10] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in Proceedings of the Third ACM International Conference on Web Search and Data Mining, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 251–260.

[11] C. Chelmis and V. K. Prasanna, "Social networking analysis: A state of the art and the effect of semantics," in Proc. ieee third international conference and 2011 ieee third international conference social computing (socialcom) Privacy, security, risk and trust (passat), 2011, pp. 531–536.

[12] L. O. Alves, M. Maciel, L. Ponciano, and A. Brito, "Assessing the impact of the social network on marking photos as favorites in flickr," in Proceedings of the 18th Brazilian Symposium on Multimedia and the Web, ser. WebMedia '12. New York, NY, USA: ACM, 2012, pp. 79–82.

[13] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in Proceedings of the 19th ACM international conference on Information and knowledge management, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 759–768.

[14] S. Chandra, L. Khan, and F. B. Muhaya, "Estimating twitter user location using social interactions–a content based approach," in Proc. ieee third international conference and 2011 ieee third international conference social computing (socialcom) Privacy, security, risk and trust (passat), 2011, pp. 838–843.

[15] H. wen Chang, D. Lee, M. Eltaher, and J. Lee, "@phillies tweeting from philly? predicting twitter user locations with spatial word usage," in Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, Aug 2012, pp. 111–118.

[16] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: Unified and discriminative influence model for inferring home locations," in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 1023–1031.

[17] L. Chen and A. Roy, "Event detection from flickr data through wavelet-based spatial analysis," in Proceedings of the 18th ACM Conference on Information and Knowledge Management, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 523–532.

[18] C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu, "Bringing order to your photos: Event-driven classification of flickr images based on social knowledge," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 189–198.

[19] C.-H. Lee, "Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams," Expert Systems with Applications, vol. 39, no. 18, 2012, pp. 13338 – 13356.

[20] H. Becker, M. Naaman, and L. Gravano, "Event identification in social media," in Proceedings of the 12th International Workshop on the Web and Databases, 2009.

[21] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu, "Mining social emotions from affective text," vol. 24, no. 9, 2012, pp. 1658–1670.

[22] M. Yassine and H. Hajj, "A framework for emotion mining from text in online social networks," in Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, dec. 2010, pp. 1136 –1142.

[23] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in Proceedings of the 3rd international workshop on Search and mining user-generated contents, ser. SMUC '11. New York, NY, USA: ACM, 2011, pp. 37–44.

[24] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1301–1309.

[25] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents, ser. SMUC '10. New York, NY, USA: ACM, 2010, pp. 37–44.

[26] K. Filippova, "User demographics and language in an implicit social network," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, ser. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 1478–1488.

[27] J. Hays and A. A. Efros, “Im2gps: estimating geographic information from a single image,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2008, 2008, pp. 1–8.

[28] J. Yu, X. Jin, J. Han, and J. Luo, “Mining personal image collection for social group suggestion,” in Proc. IEEE Int. Conf. Data Mining Workshops ICDMW ’09, 2009, pp. 202–207.

[29] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam, “Multimedia features for click prediction of new ads in display advertising,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD ’12. New York, NY, USA: ACM, 2012, pp. 777–785.

[30] P. Lovato, A. Perina, N. Sebe, O. Zandonà, A. Montagnini, M. Bicego, and M. Cristani, “Tell me what you like and i’ll tell you what you are: discriminating visual preferences on flickr data,” in Proceedings of the 11th Asian conference on Computer Vision - Volume Part I, ser. ACCV’12. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 45–56.

[31] A. Gallagher, D. Joshi, J. Yu, and J. Luo, “Geo-location inference from image content and user tags,” in Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops CVPR Workshops 2009, 2009, pp. 55–62.

[32] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua, “Research and applications on georeferenced multimedia: a survey,” Multimedia Tools Appl., vol. 51, no. 1, Jan. 2011, pp. 77–98.

[33] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, “How flickr helps us make sense of the world: context and content in community-contributed media collections,” in Proceedings of the 15th international conference on Multimedia, ser. MULTIMEDIA ’07. New York, NY, USA: ACM, 2007, pp. 631–640.

[34] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, “Mapping the world’s photos,” in Proceedings of the 18th international conference on World wide web, ser. WWW ’09. New York, NY, USA: ACM, 2009, pp. 761–770.

[35] C. Marlow, M. Naaman, D. Boyd, and M. Davis, “Ht06, tagging paper, taxonomy, flickr, academic article, to read,” in Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, ser. HYPERTEXT ’06. New York, NY, USA: ACM, 2006, pp. 31–40.

[36] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, “To search or to label?: Predicting the performance of search-based automatic image classifiers,” in Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, ser. MIR ’06. New York, NY, USA: ACM, 2006, pp. 249–258.

[37] M. Ames and M. Naaman, “Why we tag: Motivations for annotation in mobile and online media,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI ’07. New York, NY, USA: ACM, 2007, pp. 971–980.

[38] L. Wu, S. C. Hoi, R. Jin, J. Zhu, and N. Yu, “Distance metric learning from uncertain side information with application to automated photo tagging,” in Proceedings of the 17th ACM International Conference on Multimedia, ser. MM ’09. New York, NY, USA: ACM, 2009, pp. 135–144.

[39] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He, “Mining social images with distance metric learning for automated image tagging,” in Proceedings of the fourth ACM international conference on Web search and data mining, ser. WSDM ’11. New York, NY, USA: ACM, 2011, pp. 197–206.

[40] M. Wang, X. Zhou, and T.-S. Chua, “Automatic image annotation via local multi-label classification,” in Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, ser. CIVR ’08. New York, NY, USA: ACM, 2008, pp. 17–26.

[41] Y. Yang, Y. Gao, H. Zhang, J. Shao, and T.-S. Chua, “Image tagging with social assistance,” in Proceedings of International Conference on Multimedia Retrieval, ser. ICMR ’14. New York, NY, USA: ACM, 2014, pp. 81:81–81:88.

[42] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 3 Mar 2007, pp. 394–410,.

[43] W. Zhang, Z. Qin, and T. Wan, “Semi-automatic image annotation using sparse coding,” in Machine Learning and Cybernetics (ICMLC), 2012 International Conference on, vol. 2, Jul 2012, pp. 720–724.

[44] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua, “Inferring semantic concepts from community-contributed images and noisy tags,” in Proceedings of the 17th ACM International Conference on Multimedia, ser. MM ’09. New York, NY, USA: ACM, 2009, pp. 223–232.

[45] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, and G. Serra, “Tag suggestion and localization in user-generated videos based on social knowledge,” in Proceedings

of Second ACM SIGMM Workshop on Social Media, ser. WSM ’10. New York, NY, USA: ACM, 2010, pp. 3–8.

[46] I. Bartolini, M. Patella, and C. Romani, “Shiatsu: Semantic-based hierarchical automatic tagging of videos by segmentation using cuts,” in Proceedings of the 3rd International Workshop on Automated Information Extraction in Media Production, ser. AIEMPro ’10. New York, NY, USA: ACM, 2010, pp. 57–62.

[47] S. Siersdorfer, J. San Pedro, and M. Sanderson, “Automatic video tagging using content redundancy,” in Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’09. New York, NY, USA: ACM, 2009, pp. 395–402.

[48] J. S. Pedro, S. Siersdorfer, and M. Sanderson, “Content redundancy in youtube and its application to video tagging,” ACM Trans. Inf. Syst., vol. 29, no. 3, Jul. 2011, pp. 13:1–13:31.

[49] C. D. Manning and H. Schutze, Foundations of statistical natural language processing. MIT press, 1999.

[50] X. Chen, M. Vorvoreanu, and K. Madhavan, “Mining social media data for understanding students’ learning experiences,” Learning Technologies, IEEE Transactions on, vol. 7, no. 3, Jul 2014, pp. 246–259.

[51] T. Yang, D. Lee, and S. Yan, “Steeler nation, 12th man, and boo birds: Classifying twitter user interests using time series,” in Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, Aug 2013, pp. 684–691.

[52] G.-J. Qi, C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, “Exploring context and content links in social media: A latent space method,” vol. 34, no. 5, 2012, pp. 850–862.

[53] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “Liblinear: A library for large linear classification,” J. Mach. Learn. Res., vol. 9, Jun. 2008, pp. 1871–1874.

[54] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[55] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang, “World explorer: visualizing aggregate data from unstructured text in geo-referenced collections,” in Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ser. JCDL ’07. New York, NY, USA: ACM, 2007, pp. 1–10.

[56] J. Sun, S. Papadimitriou, C. Lin, N. Cao, S. Liu, and W. Qian, “Multivis: Content-based social network exploration through multi-way visual analysis,” in Proc. SDM, vol. 9, 2009, pp. 1063–1074.

[57] H. Becker, M. Naaman, and L. Gravano, “Learning similarity metrics for event identification in social media,” in Proceedings of the third ACM international conference on Web search and data mining, ser. WSDM ’10. New York, NY, USA: ACM, 2010, pp. 291–300.

[58] A. Strehl and J. Ghosh, “Cluster ensembles — a knowledge reuse framework for combining multiple partitions,” J. Mach. Learn. Res., vol. 3, Mar. 2003, pp. 583–617.

[59] A. Vakali, “Evolving social data mining and affective analysis methodologies, framework and applications,” in Proceedings of the 16th International Database

Engineering & Applications Sysmposium, ser. IDEAS ’12. New York, NY, USA: ACM, 2012, pp. 1–7.

[60] P. Heymann, D. Ramage, and H. Garcia-Molina, “Social tag prediction,” in Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’08. New York, NY, USA: ACM, 2008, pp. 531–538.

[61] B. Sigurbjornsson and R. van Zwol, “Flickr tag recommendation based on collective knowledge,” in Proceedings of the 17th International Conference on World Wide Web, ser. WWW ’08. New York, NY, USA: ACM, 2008, pp. 327–336.

[62] A. Popescu, G. Grefenstette et al., “Mining user home location and gender from flickr tags.” in ICWSM, 2010.

[63] H. Zhang, “The optimality of naive bayes,” A A, vol. 1, no. 2, p. 3, 2004.

[64] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[65] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, May 2011, pp. 27:1–27:27.

[66] N. Cristianini, J. Shawe-Taylor, and H. Lodhi, “Latent semantic kernels,” J. Intell. Inf. Syst., vol. 18, no. 2-3, Mar. 2002, pp. 127–152.

[67] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in In Workshop on Statistical Learning in Computer Vision, ECCV, 2004, pp. 1–22.

[68] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Computer Vision and Pattern

Recognition, 2006 IEEE Computer Society Conference on, vol. 2, 2006, pp. 2169–2178.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python ," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[70] P. Serdyukov, V. Murdock, and R. van Zwol, "Placing flickr photos on a map," in Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 484–491.

[71] S. Kinsella, V. Murdock, and N. O'Hare, ""i'm eating a sandwich in glasgow": Modeling locations with tweets," in Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, ser. SMUC '11. New York, NY, USA: ACM, 2011, pp. 61–68.

[72] O. Van Laere, S. Schockaert, and B. Dhoedt, "Finding locations of flickr resources using language models and similarity search," in Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ser. ICMR '11. New York, NY, USA: ACM, 2011, pp. 48:1–48:8.

[73] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," ACM Transactions on Information Systems (TOIS), vol. 22, no. 2, 2004, pp. 179–214.